ABSTRACT

HOVEY, KYLE A. Analysis and Control of Service Systems with Non-Stationary Demand. (Under the direction of Yunan Liu).

Service systems such as call-centers and emergency departments often experience demand arriving to the system at varying levels of intensity throughout their operating periods. Without reflective time-varying service capacities to accommodate this demand, customer experience in the system (e.g., waiting times) can vary greatly. This presents an operational challenge in service systems that specify constant quality of service (QoS) levels to ensure standardized customer satisfaction. To provide effective system stabilization policies, this dissertation studies representative stochastic models and establishes appropriate methodologies to achieve desired QoS results. This work will develop models and solution methods that will (a) capture system performance and (b) provide effective control policies to aid in system management.

The solution approach followed in this dissertation is detailed as follows. To conduct operational analysis, the service systems of interest are modeled as representative stochastic queuing systems. However, the inclusion of realistic model features (e.g., time-varying demand, time-varying service capacity levels, non-exponential service, patience, and abandonment times) typically renders exact mathematical analysis of stochastic system performance impossible. Therefore, the fluid and diffusion limiting models are developed to approximate the original queueing system. These limiting models provide tractable formulas that approximate the original system behavior and are used to develop potential control policies for the original system.

This dissertation consists of two pieces of work, each presented in one chapter. The first chapter (Chapter 2) proposes a new time-varying staffing rule and time-varying dynamic scheduling policy for the multi-class model, having time-varying arrivals, customer abandonment, and class-dependent service, abandonment, and arrival rates. The proposed dynamic scheduling rule prioritizes customers based on their elapsed delays with a time-varying class-dependent prioritization regulator.

Under the new staffing and scheduling policies, many-server heavy-traffic functional central limit theorem for various quantities of interest are established. By proving a state-space-collapse result for the waiting time processes, it is shown that the multiclass delay functions reduce to a simple one-dimensional process (called the frontier process). Because service rates are allowed to be class dependent, the frontier process uniquely solves a stochastic Volterra equation.

Based on the state-space-collapse and further analysis of this frontier process, we obtain the desired analytic control functions for our time-varying staffing and scheduling policies; The computation of these control functions relies on the first and second moment of the limiting frontier process for which we develop efficient algorithms. We prove that they asymptotically achieve service level differentiation for all classes at customized service targets.
Important special cases of model inputs are considered to gain useful insights of the staffing and scheduling policies. Extensive simulation experiments are conducted to substantiate the effectiveness and robustness of the results.

The second chapter (Chapter 3) develops performance analysis approximations and performance stabilization policies for a battery swapping station (BSS) serving electric vehicles (EV) owners. A BSS stores and charges EV batteries that are exchanged with depleted batteries upon a customer arrival to the system. A novel stochastic queuing model is developed to dynamically capture the continuous charge levels of all batteries in inventory in continuous time. Charge levels are treated as a function of time a battery has spent in inventory. Several model extensions are developed, each reflecting a realistic feature of service system operation such as non-linear charging and residual battery charges.

Fluid model approximations are developed for each stochastic model. System dynamics are described by establishing differential equations for battery age and charge levels. For all cases, fluid system performance processes are established to describe key quality indicators of the system. Key indicators studied include best available charge, total charge in the system, and quantity of batteries charged to given level. Monte Carlo simulations show that the fluid model approximations are effective for capturing the behavior of interest in varying model settings.

A centralized, time-varying charge rate control policy is proposed to manage BSS inventory levels. A fluid model approximation algorithm is developed to find a charge policy that will stabilize the battery charge level received by an incoming customer. In addition, a simulation-based reinforcement learning (SBRL) algorithm is created to augment the fluid method. The SBRL method proves to be robust to several model features, such as residual charge and system scale. Numerical examples were conducted to both test the proposed algorithms ability to stabilize performance, as well as, explore the sensitivity of the SBRL algorithm.
Analysis and Control of Service Systems with Non-Stationary Demand

by
Kyle A Hovey

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Operations Research

Raleigh, North Carolina
2019

APPROVED BY:

Michael Kay  
Russell King

Osman Ozaltin  
Yunan Liu  
Chair of Advisory Committee
DEDICATION

To my family, 444, 373, and 375.
ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Yunan Liu for his excellent mentoring and unwavering support through my PhD studies. I would also like to thank my committee members Dr. Michael Kay, Dr. Osman Ozaltin, and Dr. Russell King. They have provided me invaluable guidance throughout my time at NC State in my work, my teaching, and my path moving forward.

I would also like to acknowledge the countless hours that my friends have dedicated to both my academic success and happiness over the course of this program. Each one of them has forever impacted my life and I could not have completed this journey without them. To my family, I owe the biggest debt. Their absolute love and faith in me has truly made me the best student, teacher, friend, and human I could possibly hope to be. I am blessed to have my sisters Michelle and Kimberly make up two parts of the most amazing sibling trio. They have been constant and endless sources of joy in my life. Lastly, I would like to acknowledge my parents, Gene and Candy. Words will never be able to express or describe how much of what is good in my life is because of them. Every single day of my life they have had my back and selflessly given me everything so that I could be where I am today.
## TABLE OF CONTENTS

**LIST OF TABLES** ......................................................................................... vii

**LIST OF FIGURES** ....................................................................................... viii

**Chapter 1  Introduction** ............................................................................... 1
  1.1 Time-Varying Demand in Service Systems .............................................. 1
  1.1.1 Time-Varying Staffing to Achieve Performance Stabilization ............. 2
  1.2 Customer Abandonment ......................................................................... 5
  1.3 Non-Markovian Probability Structure ...................................................... 6
  1.4 Multiclass Models .................................................................................... 8
  1.5 Many-Server Heavy-Traffic Approximations ......................................... 9
    1.5.1 Heavy-Traffic Limits ...................................................................... 9
    1.5.2 MSHT Fluid Models .................................................................... 10
    1.5.3 Diffusion Limits ........................................................................ 11
  1.6 State Space Collapse ............................................................................ 12
  1.7 Numerical Studies .................................................................................. 13
    1.7.1 Numerical Approximation of Continuous Functions ....................... 13
    1.7.2 Discrete Event Simulation .......................................................... 14

**Chapter 2  Staffing and Scheduling to Differentiate Service in Time-Varying Multiclass Service Systems** ......................................................... 17
  2.1 Introduction ............................................................................................ 17
  2.1.1 Literature ........................................................................................ 19
  2.1.2 Contributions .................................................................................. 20
  2.2 Problem Formulation and Proposed Solutions ....................................... 21
    2.2.1 A Multiclass V model .................................................................... 21
    2.2.2 A Time-Varying Square-Root Staffing Rule .................................. 23
    2.2.3 A Time-Varying Dynamic Prioritization Scheduling Rule .......... 25
  2.3 Asymptotic Service-Level Differentiation .............................................. 26
    2.3.1 Many-Server FCLT Limits under TV-SRS and TV-DPS ................. 26
    2.3.2 Asymptotic Service Differentiation and Stabilization ................. 32
    2.3.3 The Case of Class-Independent Service Rate ................................ 34
  2.4 Numerical Studies .................................................................................. 36
    2.4.1 A Two-Class Base Model ............................................................. 36
    2.4.2 Other Cases .................................................................................. 41
    2.4.3 Implementation Details .................................................................. 46
    2.4.4 Additional Numerical Examples ................................................. 46
    2.4.5 Staffing and Scheduling to Differentiate the Mean PWT ................ 49
  2.5 Proofs .................................................................................................... 50
    2.5.1 Proof of Theorem 2.3.1 .................................................................. 51
    2.5.2 Proof of Proposition 1 ................................................................... 60
    2.5.3 Proof of Proposition 2 ................................................................... 65
Chapter 3  Modeling and Analysis of Battery Swapping Station for Electric Vehicles

3.1  Introduction 70

3.1.1  Background and Motivation of Battery Swapping Stations 70
3.1.2  Management of BSS Charge Inventory 71
3.1.3  Contributions 72
3.1.4  Literature Review 73
3.1.5  Chapter Organization 74

3.2  Stochastic Model for a BSS 75

3.2.1  Empirical Evidence of Realistic Features 75
3.2.2  Base Model Formulation 77
3.2.3  System Processes 79
3.2.4  Key Performance Metrics 82

3.3  The BSS Fluid Model 84

3.3.1  Fluid Model Dynamics 84
3.3.2  Performance Functions 86
3.3.3  General Charging Rates 89
3.3.4  Residual Charge Extension 90
3.3.5  Engineering Confirmation via Simulations 96
3.3.6  FWLLN for BSS under General Charging Function 100

3.4  Achieving Time-Stable Service Quality 102

3.4.1  Service Quality Control of a BSS 103
3.4.2  Fluid Model Algorithm for Performance Stabilization of a BSS with No Residual Charge 104
3.4.3  A SBRL Algorithm for Performance Stabilization 110
3.4.4  Numerical Examples 114

3.5  Concluding Remarks 125

Chapter 4  Conclusions and Future Work 128

4.1  Dissertation Summary and Contributions 128
4.2  Future Work 129

References 132

Appendices 143

Appendix A  Appendix for Chapter 2 144
A.1  Multiclass Service System without Abandonment 144
A.1.1  Many-Server Heavy Traffic Functional Central Limit Theorem 144
A.1.2  Distribution of Frontier Process 146
A.1.3  Asymptotically Unique Control Functions 147
A.1.4  Differentiation of the mean PWT 148
A.1.5  Class Independent Service Rate 150
A.1.6  Constant Arrival 151
A.1.7 Numerical Results ......................................................... 152
A.1.8 Conclusions ............................................................... 159

Appendix B Appendix for Chapter 3 ........................................ 162
B.1 Proof of Theorem 3.3.1 .................................................... 162
B.1.1 FWLLN for $W_n$ ....................................................... 163
B.1.2 FWLLN for Key Performance Indicators ......................... 164
B.2 Proposed Fluid Model Functions for Stabilization of FHCL with Residual Charge .................................................. 165
LIST OF TABLES

Table 1.1  Comparison of performance metrics between an \( M/M/50+M \) and \( M/M/50 \) model with call-center parameters of 48 call per minute, 1 minute average service time, 2 minute average patience [33]. .................................................. 6

Table 2.1  A two-class base case: Average, max, and min of (simulated) TPoD’s and relative differences to their target levels, using floored, rounded, and ceiled versions of the TV-SRS formula. .................................................. 39

Table 2.2  Five Class Model: Class specific parameters and QoS target levels for the TPoD stabilization model. .................................................. 45

Table 2.3  Five Class Model: Class specific parameters and QoS target levels for the PWT stabilization model .................................................. 50
LIST OF FIGURES

Figure 1.1 (a) Hourly call rates for a bank call center [18]; (b) Averaged hourly arrival counts of customers to a battery swapping station in Beijing, China ................................. 2
Figure 1.2 The DIS approximation for a single queue with delay target of \( w \) ....................... 4
Figure 1.3 An example of a continuous staffing function discretized using three methods rounding, ceiling, and flooring ....................................................... 5
Figure 1.4 Histogram of log(service time) of a call-center [18] ........................................ 7
Figure 1.5 A multiclass V model with time-varying arrivals .................................................. 22
Figure 1.6 Road map of our approach .................................................................................. 21
Figure 1.7 An example of a continuous staffing function discretized using three methods round-
ing, ceiling, and flooring ................................................................................. 5
Figure 1.8 Histogram of log(service time) of a call-center [18] ........................................ 7
Figure 1.9 Mixed Target Delay (\( w_1 = 0.5, w_2 = 2 \)): Comparison of Estimated Class Time-Dependent TPoD, \( P(V_i(t) > w_i) \), from 5000 Independent Simulation Runs to Class Target TPoD QoS Levels, \( \alpha_i \) ................................................................. 44
Figure 1.10 A five-class based model: (i) Computed control functions \( m(t), c(t) \), and \( \kappa_i(t) \) for \( i = 1, \ldots, 5 \) (left), (ii) Simulation comparisons for TPoD \( P(V_i(t) > w_i) \), \( i = 1, \ldots, 5 \) (right), with \( n = 50 \), input and QoS parameters given in Table 2.3, and 5000 samples. ................................................................................. 45
Figure 1.11 Simulation comparison for a two-class base case: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD \( P(V_i(t) > w_i) \) (middle panel); and (iii) time-varying ASL and MSL staffing levels having \( \Delta_s = 0.5 \) (bottom panel), with \( w_1 = 0.5, w_2 = 1, \alpha_1 = 0.2, \alpha_2 = 0.8 \), and 5000 independent runs. ......................................................... 47
Figure 3.28 Convergence of Charge Rate Function under the SBRL Algorithm for the Lognormal Residual Case .......................................................... 121

Figure 3.29 Simulation Results of SBRL Algorithm Performance Stabilization of HCL, TCL, and Proportion Charged at Least 50%; Uniform Residual Case .................. 122

Figure 3.30 Convergence of HCL under the SBRL Algorithm for the Uniform Residual Case .......................................................... 123

Figure 3.31 Convergence of Charge Rate Function under the SBRL Algorithm for the Uniform Residual Case .......................................................... 123

Figure 3.32 Simulation Results for SBRL Algorithm Performance Stabilization of HCL, TCL, and Proportion Charged at Least 50% with $L = 0.95$ .................................. 124

Figure 3.33 Simulation Results for SBRL Algorithm Performance Stabilization of HCL, TCL, and Proportion Charged at Least 50% with Initial Candidate Value $\mu^0 = 1.5$ .... 125

Figure 3.34 Convergence of HCL under the SBRL Algorithm with Initial Candidate Value $\mu^0 = 1.5$ .......................................................... 126

Figure 3.35 Convergence of Charge Rate Function under the SBRL Algorithm with Initial Candidate Value $\mu^0 = 1.5$ .......................................................... 126

Figure A.1 Simulation results for a two-class base case with no abandonment: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $P(V_i(t) > w_i)$ (middle panel); and (iii) time-varying staffing level (bottom panel), with $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 1000 independent runs. .................................................. 153

Figure A.2 Simulation comparison for a two-class base case with no abandonment: (i) arrival rates (top panel); (ii) simulated class-dependent delay $E[w_i]$ (top middle panel); (iii) simulated class queue length $E[Q_i]$ (bottom middle panel); and (iv) simulated probability of empty class queue $P(Q_i(t) = 0)$ (bottom panel), with $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 1000 independent runs. .................................................. 154

Figure A.3 Simulation results for a two-class base case with no abandonment and $n = 200$: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $P(V_i(t) > w_i)$ (middle panel); and (iii) time-varying staffing level (bottom panel), with $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 1000 independent runs. .................................................. 155

Figure A.4 Simulation comparison for a two-class base case with no abandonment and $n = 200$: (i) arrival rates (top panel); (ii) simulated class-dependent delay $E[w_i]$ (top middle panel); (iii) simulated class queue length $E[Q_i]$ (bottom middle panel); and (iv) simulated probability of empty class queue $P(Q_i(t) = 0)$ (bottom panel), with $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 1000 independent runs. .................................................. 156

Figure A.5 Simulation comparison for a two-class base case with no abandonment: (i) arrival rates (top panel); (ii) simulated class-dependent delay $E[w_i]$ (top middle panel); (iii) simulated class queue length $E[Q_i]$ (bottom middle panel); and (iv) simulated probability of empty class queue $P(Q_i(t) = 0)$ (bottom panel), with $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.5$, $\alpha_2 = 0.5$, and 1000 independent runs. .................................................. 158

Figure A.6 Simulation results for a two-class base case with no abandonment: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $P(V_i(t) > w_i)$ (middle panel); and (iii) time-varying staffing level (bottom panel), with $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 1000 independent runs. .................................................. 159

Figure A.7 Computed control functions for a two-class base-case with no abandonment and constant arrival rates: $m(t), c(t), \kappa_i(t)$ and $\sigma(t)$, $i = 1, 2$. .................................................. 160
Chapter 1

Introduction

This work is motivated by the desire to both model performance and effectively manage service systems with time-varying demands, primarily through the use of fluid and diffusion approximation models. These service systems can include emergency departments, call centers, and vehicle service stations. The goal of controlling a service system is often to achieve a desired system performance. These performance levels can be comprised of either the quality of service a customer receives, the costs associated with operating the system, or a mixture of both. Achieving these performance levels in the face of time-varying demand often requires a time-varying control of the system, i.e., time-varying staffing levels and scheduling policy. This dissertation develops system performance functions for both a general multiclass service system and an electric-vehicle service station under time-varying demand. We accomplish our goal through (i) the construction of representative stochastic queueing models, (ii) the establishment of appropriate diffusion and fluid limits to approximate stochastic system dynamics, and (iii) the treatment of these models and limits to approximate desired results in the original system. This works studies the fluid and diffusion approximations as these methods can effectively treat not only the time-varying nature of the developed queueing models, but other realistic features such as abandonment and non-Markovian system behavior. In addition, we provide a reinforcement learning method for the electric vehicle service station that exploits stochastic system simulation to augment the fluid approximations of small scale systems. §1.1-1.4 will review these realistic model features, §1.5-1.6 will discuss the mathematical framework for asymptotic approximations, and §1.7 reviews stochastic system simulation.

1.1 Time-Varying Demand in Service Systems

Service systems with customer arrivals often experience varying levels of intensity and resulting congestion throughout the day. This temporal fluctuation can be captured through a time-varying arrival rate of customers. Figure 1.1 (a) shows the average hourly incoming call rates for a bank are highly non-stationary. There is a sharp increase in arrivals over the hours of 7-10am with an almost immediate
Figure 1.1: (a) Hourly call rates for a bank call center [18]; (b) Averaged hourly arrival counts of customers to a battery swapping station in Beijing, China
decrease until 12. However, for the battery swapping station (BSS) serving a taxi company in Figure 1.1 (b), the biggest decrease in arrivals occurs during mid-afternoon with a following peak at 8pm. Further examples of time-varying arrival rates to service systems can be found in [38, 91, 105]. A natural question that could arise is how do system managers administer in the face of time varying demand to achieve desired system performance.

1.1.1 Time-Varying Staffing to Achieve Performance Stabilization

Of the examples presented in Figure 1.1, the arrival pattern of the bank call-center suggests that the bank should bolster the service capacity of the call-center during the relative peaks during the mid-morning and mid-afternoon time periods. This should be done in an effort to ensure customer experiences such as wait times are maintained. Quality of service (QoS) standards for a system are often defined in terms of constant targets even in time-varying systems. Call-centers often employ the so-called 80/20 rule that stipulates 80% of customers should be answered within 20 seconds [76]. As we have seen, call-center arrivals can be highly time-varying and achieving the 80/20 performance target would require different levels of service resources (number of call-center representatives) throughout the day. In this proposal, we focus on the class of problems that are concerned with achieving a stabilized system performance (i.e. stabilized queue lengths, waiting times, etc.) in the face of time-varying demand. This stabilization is often achieved through the specification of time-varying service capacity.

Approximations of time-varying staffing systems

Capturing the dynamics of a service system with time-varying arrivals is often difficult. Stochastic models often become impossible to solve directly for desired performance metrics, thus leading to several approximation methods for queueing models. Defraeye et al. provides a review of staffing under non-
stationary demand [25]. We review several approximation techniques.

**Point-wise stationary approximation.** Green et al. [37] proposed the pointwise stationary approximation (PSA) which segments the queuing system with time-varying arrivals into a series of stationary queues indexed by time with the time-varying rate fixed as an stationary arrival rate for each queue. Performance measures of the original time-varying system are approximated as the weighted average of performance measures taken from the series of stationary queues. This approximation performs well as the individual service rate increases with the instantaneous traffic intensities remaining unchanged, with the limiting behavior proven to be asymptotically correct by Whitt in [101]. An additional appeal to this method is that it lends itself naturally to modeling time-varying customer behavior as well (i.e. patience and service times that depend on time-of-day). We write the approximation as

\[ M_t/M_t/s_t + M_t \approx (M/M/s + M)_t. \]

However, PSA breaks down in effectiveness as the service rates become small in comparison with the arrival rate. and other approximation methods are needed to determine performance measures [69].

**Offered load and delayed infinite server.** To circumvent complicated and often intractable staffing calculations, the offered load (OL) approach makes use of the more tractable infinite server (IS) queue. That is all customers who arrive to the system are immediately taken into service. The time-varying number of busy-servers in the infinite server system is calculated to determine the load on the system. For instance, the \( M_t/G/s_t + G \) model is analytically intractable but the \( M_t/G/\infty \) has tractable equations for busy servers. By staffing the system according to the OL approximation, the system can effectively match the service resources required by incoming demand. See [74] for an application of OL to call-centers.

In an effort to stabilize the tail probability of delay, the modified offered load (MOL) was proposed in [29] and [49] as a further refinement to the simple OL method. It combined a pointwise stationary approximation approach with OL results to create a sequence of stationary queues with arrival rates to each stationary queue determined by the offered load over the expected service time, which arises from Little’s Law. Both the OL and MOL models are utilized to better match incoming demand with service capacity, thus aiming at having no waiting customers and not considering customer abandonment.

The delayed infinite server (DIS) MOL proposed in [64] attempted to stabilize customer abandonment by employing two IS queues in series, with the first queue representing the waiting room and the second representing those actually in service. The first IS queue has deterministic service times that were set at the target abandonment times. The resulting offered load in the second IS queue was then used in the refine MOL technique to create a staffing function recommendation that effectively stabilized the tail probability of delay. A graphical representation of the DIS approach is given in Figure 1.2. A
The DIS approximation for a single queue with delay target of $w$

target wait time of $w$ is imposed as the deterministic service time of the waiting room IS. To determine the approximate staffing capacity needed to achieve this waiting time, the number of busy servers in the service facility IS is then calculated.

**Large scale asymptotic analysis.** In recent years, groundbreaking work has been conducted on using asymptotic analysis to analyze time-varying queues [67]. Generally, the method revolves around studying the limiting dynamics of stochastic queueing models. The method followed in this work first proposes a control on the system (staffing, scheduling, arrival routing, etc...) for the stochastic model. It is then shown to achieve the desired system performance stabilization in the limit. These limits are often tractable and can be deterministic. The proposed control is then shown to be adequate at stabilizing the original queue. We discuss this method in more detail in Chapter 2, where we consider a joint staffing and scheduling control to achieve stabilized performance, shown by asymptotic results.

**Time-varying staffing implementation**

We highlight that we are providing the total quantity of service resources required to achieve stabilized performance, not the method in which the staffing of required resources is completed. Similar to the works in [29] and [38], servers are treated as identical interchangeable entities with no limitations on the amount of time they remain in system. We answer the question of how many servers are needed at a specified time throughout the day. This is an important distinction from the widely studied personnel scheduling problem which attempts to answer the question of how to schedule individual servers fairly and effectively to meet total service load. See [98] for a comprehensive overview of this type of problem.

The staffing functions presented in this work are continuous functions, expressed in terms of continuous arrival rates. Continuous staffing presents two challenges in most service systems where humans are the service resource. The first is that fractional values for service resource are evidently not appealing to system managers who need to staff an integer number of servers. Recognizing this importance, all staffing functions were rounded up to the nearest integer, (see Figure 1.3 for a graphical depiction of several methods of discretizing). In systems where there are a large number of servers, the effect of this discretization diminishes. The effect of adding or removing a server when there are currently 100
servers will have a smaller impact than in the same system with only 10 servers. See §2.4 for further discussion.

The second challenge is the temporal flexibility with which managers can staff resources. Workers are often scheduled in shifts for a designated period of time, thus rendering the staffing resource level constant throughout these shifts. The flexibility or inflexibility of the staffing throughout the day depends on the service system considered. Call-centers have been modeled with a flexibility of 30 minute staffing intervals [82], while nurse scheduling tends to be modeled on a daily basis [75] or in shifts that have 8 or 12 hours [104]. Continuous staffing is more effective when the variability required is on the same order of staffing flexibility allowed by the system. See §2.4 for further discussion.

1.2 Customer Abandonment

Customer patience is an important concept to incorporate when modeling service systems. If a customer does not receive service before their intrinsic patience time has elapsed, then a customer will abandon the system without being seen. Abandonment is an important feature to capture as it can indicate the quality of experience received by a customer and increase the accuracy of the model. From a quality of service perspective, abandonment is often treated as a negative reflection on a system’s capacity to meet demand. In call centers this is realized as customers hanging up before being helped by a service representative after they have been on hold for too long. Generally, call center managers strive to minimize the number of abandonments [3]. Walk-in patients to a hospital will abandon the system if it is taking too long to see a doctor in a phenomena know as “left without being seen” (LWBS) [105]. From a model-building perspective, the inclusion of abandonments can drastically alter the system performance and is an important feature when attempting to create practical models. See Garnett et al. [33, 34] for a discussion of model inclusion of abandonment in call-centers. Abandonment from queue has a natural load correcting effect on the system (greater number of abandonments from queue leads to a more decreased queue length). We give a call-center example taken from [33] highlighting a comparison of a
Table 1.1: Comparison of performance metrics between an $M/M/50 + M$ and $M/M/50$ model with call-center parameters of 48 call per minute, 1 minute average service time, 2 minute average patience [33].

<table>
<thead>
<tr>
<th></th>
<th>$M/M/N$</th>
<th>$M/M/N + M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction abandoning</td>
<td>—</td>
<td>3.1%</td>
</tr>
<tr>
<td>Average speed of answer</td>
<td>20.8 sec.</td>
<td>3.6 sec.</td>
</tr>
<tr>
<td>Waiting time’s 90th percentile</td>
<td>58.1 sec.</td>
<td>12.5 sec.</td>
</tr>
<tr>
<td>Average queue percentile</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>Agents’ utilization</td>
<td>96%</td>
<td>93%</td>
</tr>
</tbody>
</table>

$M/M/50$ model with abandonment and without. The call-center considered had 50 agents, 48 calls per minute arriving, 1 minute average service times, and 2 minute average patience times. Shown in Table 1.1 are the numerical results comparing the average waiting times and queue lengths.

As seen in Table 1.1, with only a 3% proportion of the customers abandoning the system, the abandonment model achieved significantly decreased waiting times and queue lengths. Customer were expected to have their calls answered within approximately 4 seconds as opposed to 21 seconds in the non-abandonment model. Another important consideration is abandonments from service systems can also lead to customer retrials as they have not received their desired service. In the call center, this is known as the reconnect effect and can impact the true demand on a call center, see [26].

To model customer abandonment behavior in a queueing system, we typically assume each customer has an exogenous random patience time, meaning it is independent of system factors (i.e. current queue lengths or number of busy servers). This patience time is the total amount of time this particular customer is willing to spend waiting in queue before entering service. Using this patience time, all customers upon entering the system have a corresponding abandonment time at which they will depart the waiting queue and exit the system. If the time of a customer to enter service is less than the abandonment time, the customer will enter service and the abandonment behavior is no longer valid. Abandonment is considered almost exclusively from the waiting queue. Customers of the same class have patience times distributed according to the same distribution. For a review on endogenous patience times that are affected by congestion levels see [9, 46].

1.3 Non-Markovian Probability Structure

-markov property- Queueing models have been well studied under the appealing mathematical Markovian structure, with the $M/M/1$ queue often the first queueing example given in textbooks. The Markov property of this model implies that customer behavior (interarrival times and service times) is distributed according to the exponential distribution. The exponential distribution has the desirable memoryless
property, which indicates that a customer’s remaining service time does not depend on the amount of time already served. The appeal of making this limiting assumption is that it gives rise to continuous-time Markov chain (CTMC) stochastic processes which are tractable and well studied. Even in the case of non-exponential distributions Markov chains can be employed if sufficient historical information is captured within the state definitions. However, the addition of extra information causes the state space of the Markov process to suffer from the so called curse of dimensionality \cite{96}. While there exist cases where the exponential distribution is an adequate empirical fit to data, see \cite{44}, it has been shown that call-center service times can be empirically better suited to distributions such as gamma and lognormal \cite{79} \cite{13} \cite{18}, an example of which is seen in Figure 1.4. Brown also shows in \cite{18} that amount a time a customer was willing to wait on the phone before hanging up (abandoning) could be shown to have a non-constant hazard-rate function, thus making the exponential distribution a poor fit. Time-varying arrivals are typically modeled by the non-homogeneous Poisson process (NHPP), which maintains a variance-to-mean ratio of 1 at all times. However, when the arrival data shows over-dispersion (ratio greater than 1) or under-dispersion (ratio less than 1), the NHPP modeling assumption loses tractability. Kim and Whitt \cite{55, 54} proposed a test to check the validity of the NHPP assumption and found that over multiple days there is significant over-dispersion of arrivals to a call-center. See \cite{13, 18} for more discussion on the modeling of general arrivals to a call-center. \cite{53} showed that arrivals to an appointment based medical clinic experienced under-dispersion due to the more deterministic scheduling process. This motivates the study of a general time-varying arrival process to model these over- and under-dispersed arrival processes, see \cite{35, 47, 61, 62} for construction and analysis of such $G_t$ arrival processes.

Figure 1.4: Histogram of log(service time) of a call-center \cite{18}
Because of the existence of non-Markovian customer behavior in many aspects of queue models, (i.e. arrivals, patience, and service times) we are motivated to study stochastic models with non-Markovian features to provide robust tools for service system modelling.

1.4 Multiclass Models

In simple queueing models, customers are typically treated as one homogeneous class that behave according to a single set of probabilistic variables, such as patience times and service times. A multiclass model refines these simple models by allowing different classes of customers to be treated in a single system according to class dependent probabilistic structures. For instance, a call center representative can be responsible for responding to emails as well as handling customer calls. To more accurately model the workload on a representative, we would need to treat the calls and emails as requiring inherently different probability structures. For instance, emails jobs arrive to the call-center at 50 emails per hour but require an average of 3 minutes to process. In contrast, call jobs arrive less frequently, 20 calls per hour but require and average of 10 minutes to address. In addition, as discussed in §1.2 customers on hold will abandon, yielding an abandonment rate of call jobs.

When different classes share a similar service resource, the management of a system becomes not only how to staff servers but how to choose jobs to route into service from all available classes. Several basic routing policies arise in an effort to achieve improved system performance. Priority queues specify an order in which different class customers should be served and polling models switch between class queues and service all customer therein until the queue is depleted [100]. To analyze performance in multiclass service systems without any priority given to any class, a global first-come-first-served (FCFS) policy can be used. This policy treats all customer classes as if they entered a single queue and are then served on a FCFS basis. This policy provides a fairness to all customers based on arrival times, but disregards class distinction. However, with the distinction of different classes, often comes a distinction in required QoS levels. For instance, the Canadian triage and acuity scale (CTAS) guideline is widely used in Canadian emergency departments to both classify incoming patients and determine the appropriate level of service required by hospital regulation for each class [27, 19]. There are five levels of classification ranging from level 1, “resuscitation”, to level 5, “non-urgent” [39]. The CTAS guideline states that “CTAS level $i$ patients need to be seen by a physician within $w_i$ minutes $100\alpha_i\%$ of the time”, with

\[
(w_1, w_2, w_3, w_4, w_5) = (0, 15, 30, 60, 120) \quad \text{and} \quad (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0.98, 0.95, 0.9, 0.85, 0.8).
\]

According to the guideline, a level 1 patient must be seen immediately 98% of time, while a level 5 patient must wait no more than 2 hours 80% of the time. This represents a drastic differentiation in service experience between the classes, in both the wait time and probability of meeting the specified
wait time target. In an overloaded emergency department, with multiple waiting customers of varying levels of acuity, meeting these target requirement requires a careful consideration on how to pick which customer is next seen by an available hospital staff. In Chapter 2, we will show that a time-varying and state-of-system dependent routing policy of customer classes into service achieves stabilized customer performance in just such a case.

## 1.5 Many-Server Heavy-Traffic Approximations

Because this work assumes both time-varying dynamics and not necessarily Markovian system behavior, exact analysis of the system becomes intractable and we rely on approximations to gain insight into the system. Because the system is time-varying we are also unable to perform steady state analysis as it does not exist for a truly time-varying system. To approximate the stochastic system, we first study the limiting behavior of the proposed queueing systems as the load on the system grows to be large. These limits are known as the heavy-traffic limits.

### 1.5.1 Heavy-Traffic Limits

The heavy-traffic limits arise from a sequence of stochastic queueing models (known as pre-limit models) that have increasingly large traffic intensities (thus heavy-traffic). The pre-limit performance functions of the model (i.e. number in queue and wait times) will typically diverge to infinity. Thus, to achieve nondegenerate limits, the output of the system must be scaled appropriately as well [1]. There are two types of common scalings yielding nondegenerate limits: conventional and many-server. This work uses the many-server heavy-traffic limit exclusively, but we review the conventional for completeness.

**Conventional heavy-traffic limit**

The conventional heavy-traffic (HT) limit holds the number of servers fixed and scales the output of the system by scaling service rates. This scaling increases the traffic intensities, loosely described as input over output, to the critical value of 1; see [102]. Kingman [56] first showed a heavy traffic limit and approximation for the $GI/GI/1$ queue that was able to approximate the steady-state waiting time for the stochastic queue as a function of the limiting steady-state waiting time and the scaled traffic intensity. The conventional heavy-traffic limit was extended to a model with multiple servers in [15]. [48] provided an important extension from steady-state convergence results to transient process results. They were able to show the convergence of the entire queue length process as function of time. However, conventional heavy-traffic limits lose appeal when approximating systems with multiple servers.
**Many-server heavy-traffic limit**

Because this work is focused on providing practical models for service systems with many servers, we turn to the many-server heavy-traffic (MSHT) limit to provide better approximations. This limit was first established by Halfin and Whitt for [43] for a $GI/M/s$ model. They proposed simultaneous scaling of the arrival rates and number of servers. For example, if we let $n$ be the number of servers and $\lambda_n > 0$ be an appropriate arrival rate scaled by $n$, then the $n^{th}$ model has $n$ servers and an $n\lambda$ arrival rate and a traffic intensity of $\rho_n \equiv \frac{\lambda_n}{n\mu}$.

MSHT limits exist in three regimes as named in [33]: quality driven (QD), efficiency driven (ED), and quality-efficiency driven (or Halfin-Whitt regime). These regimes are defined by the limiting behavior of the traffic intensity ($\rho$) as we are no longer directly controlling the traffic intensity as in the conventional heavy-traffic limit. Quality and efficiency speak to limiting behavior that yields no waiting customers and no idle servers, respectively. The QD regime has $\rho < 1$ where the limiting behavior has idle servers and no queue length, thus quality focused. The ED regime has $\rho > 1$ where there is a positive queue length and no idle servers, thus efficiency drive. The QED regime has a balanced input and output yielding both no idle servers and queue length, thus achieving both quality and efficiency. QED MSHT has been widely studied due to the tractability of the model and balanced nature of the system.

### 1.5.2 MSHT Fluid Models

A fluid model is a deterministic model describing the behavior of atoms of fluid through a queueing system, often in terms of input and output rates. These fluid models have been shown to be effective approximations of service [14], [36]. Furthermore, fluid models are mathematically linked to the original stochastic system in that they are the limiting behavior as can be shown through the functional strong law of large numbers (FSLLN). We can create a fluid representation of queueing systems by letting exogenously arriving fluid to a fluid buffer represent arriving customers. Fluid is pulled from the other end of the buffer at a certain rate and represents customers entering service. Using simple calculus and the input and output rates, we can determine the quantity of fluid in the buffer at any given time. In Chapter 3 we develop a novel fluid model approximation for a battery swapping station (BSS). Fluid models provide a continuous, deterministic approximation that is often far more tractable than its discrete, stochastic pre-limit counterpart. Let $X(t)$ be the number of customers present in a stochastic queueing model. $X(t)$ is a random process as it is a function of random arrival, patience, and service times. Let $X_n(t)$ represent the total customer present in sequence of scaled queues indexed by $n$, either through HT or MSHT scaling. Then FSLLN states that

$$\frac{X_n(t)}{n} \overset{a.s.}{\rightarrow} \bar{x}(t), \text{ as } n \to \infty$$
where $\bar{x}(t)$ is the deterministic fluid model process and $\xrightarrow{a.s.}$ indicates almost sure convergence. This means that the limiting scaled stochastic process will be $\bar{x}(t)$ with probability 1. See [102] for a further discussion of convergence. As $n$ grows to be large through heavy-traffic scaling, the stochastic fluctuations of the pre-limit process become negligible. As a result, the approximation

$$X_n(t) \approx n \cdot \bar{x}(t)$$

increases in effectiveness as the scale of the system increases or equivalently $n$ increases. This scaling is appropriate for customer quantity processes (i.e. number in queue, number of arrivals). However, processes related to customer experience (i.e. potential waiting times) are not scaled. For instance, let $W_n(t)$ be the potential waiting time of a customer arriving at time $t$ in the $n^{th}$ stochastic system and $\bar{w}(t)$ be the deterministic waiting time in the fluid model. Then we have

$$W_n(t) \xrightarrow{a.s.} \bar{w}(t), \quad \text{as } n \to \infty$$

where no corrective scaling is needed. This work focuses on the fluid models derived from the MSHT regime similar to [34] [71] [65]. [65] provides a sophisticated fluid limit for the $G_t/GI/s_t + GI$ queue.

### 1.5.3 Diffusion Limits

Diffusion processes provide a refinement to the fluid model approximation as they quantify the random fluctuation of the original stochastic process around its fluid limit. We refer to the fluid and diffusion information as first- and second-order information, respectively. This provides a refinement to the first order information as well as capturing second-order information. Diffusion processes are captured through the functional central limit theorem (FCLT). Returning to the stochastic processes defined in the previous section we have

$$\frac{X_n(t) - \bar{x}(t)}{\sqrt{n}} \Rightarrow \hat{X}(t) \quad \text{in } \mathbb{D} \quad \text{as } n \to \infty$$

where $\hat{X}(t)$ is some form of a Brownian motion and $\Rightarrow$ indicates convergence in distribution. This convergence implies that the limiting probabilistic behavior of the scaled stochastic process follows the same distribution as $\hat{X}(t)$. This convergence is defined on the function space of all right-continuous $\mathbb{R}^k$-valued functions with left limits defined on a subinterval $I$ of the real line which we denote as $\mathbb{D}$, [102]. Because the processes of interest in this thesis are non-negative quantities such as queue lengths and wait times, the sub interval $I$ considered will be $\mathbb{R}_+ \equiv [0, \infty)$. Just like for FSLLN, the potential waiting time requires careful scaling to achieve appropriate FCLT results,

$$\sqrt{n} (W_n(t) - \bar{w}(t)) \Rightarrow \hat{W}(t) \quad \text{in } \mathbb{D} \quad \text{as } n \to \infty,$$
where again we scale by $\sqrt{n}$. Note that in QED and QD regimes, the fluid level centering can be zero in the case of queue lengths and wait times. Using this result, we can now approximate the time-dependent behavior of the $n^{th}$ scaled stochastic system as

$$X_n(t) \overset{d}{=} n \cdot \bar{x}(t) + \sqrt{n}\hat{X}(t),$$

where $\overset{d}{=}$ is approximate in distribution. We can now achieve an approximation for both the second-order variance information and a defining probability distribution of the original stochastic process. Using this result, we can now approximate the time-dependent behavior of the $n^{th}$ scaled stochastic system as

$$\text{Var}[X_n(t)] \approx n \text{Var}[\hat{X}(t)], \quad P(X_n(t) > y) \approx P\left(\hat{X}(t) > \frac{y - n\bar{x}(t)}{\sqrt{n}}\right)$$

For a further review of FWLLN and FLCT under MSHT scaling limits for queues see [68] and the references within.

1.6 State Space Collapse

A critical aid in the proof of heavy-traffic limit convergence theorems for diffusion-scaled performance processes is the establishment a so-called state space collapse (SSC) result, especially in multiclass system where the dimensionality of relevant stochastics processes is often high. This SSC result is often achieved through the careful control (by means of routing and staffing customers) of the prelimit process. When the model is scaled to its HT limits, the reduction or collapse in dimensionality occurs. For a simple example, we discuss the HT limit for two parallel $M/M/1$ queues under the join-the-shortest-queue (JSQ) routing policy for arrivals, see [106] for reference. The JSQ policy states that an arriving customer will be routed to the queue that has the smallest number of current waiting customers. The HT diffusion limit for the number of customers in each queue was shown to converge to one-half the diffusion limits of a $M/M/2$ queue. Let $\hat{Q}_{n,1}(t), \hat{Q}_{n,2}(t)$, and $\hat{Q}(t)$ be the prelimit scaled diffusion process for queue 1 and 2 under JSQ and the diffusion limit for the 2 server case respectively. Then we have,

$$\left(\hat{Q}_{n,1}(t), \hat{Q}_{n,2}(t)\right) \Rightarrow \left(\frac{\hat{Q}(t)}{2}, \frac{\hat{Q}(t)}{2}\right)$$

Thus, the two-dimensional process achieved a state space collapse to the one-dimension case. Later in Chapter 2, we will show that class dependent head-of-line (HOL) waiting times under our proposed staffing and scheduling policy achieves a SSC result. Essentially, the $K$-dimensional HOL waiting time process reduces to a single dimension frontier HOL process. In [41], they were able to show that under their queue-and-idleness-ratio rule for a multiclass, multi-group server system that the class dependent
stochastic diffusion scaled processes for queue length, idle servers, and virtual waiting time had heavy-
traffic limits that depended on the single diffusion limit for total number of customers in system. This
is a powerful result as now instead of developing diffusion limits for each individual class processes, a
single limit for total number of customer in system could be established. SSC results have been studied
in both the conventional HT and MSHT regimes. We refer to [21] for an overview of SSC in HT limits.
In [16], Bramson was able to show that in the conventional HT regime, it suffices to show a similar
SSC result for the fluid model to achieve diffusion results. This result allowed advances HT limits in
multiclass servers as in [17, 72]. Armony has shown practical applications in the MSHT regime by
applying a SSC result to approximate call center dynamics [7, 8].

1.7 Numerical Studies

In consort with the asymptotic analysis conducted in this thesis, extensive numerical studies were con-
ducted to observe the performance of the stochastic models being studied. All computer simulation and
numerical analysis was done in Matlab.

1.7.1 Numerical Approximation of Continuous Functions

The performance and control functions established in this thesis are done so using the fluid and diffusion
approximations discussed in §1.5 and are therefore most often continuous functions and non-solvable.
To overcome this difficulty, numerical approximation was employed to estimate the processes over a
time-horizon \([0, T]\), where \(T = 24\) in this thesis. This was done to model a 24 hour cycle of a typical
service system. The calculated functions were captured at small intervals of size 0.001 and the result-
ing staffing and scheduling controls were stored in the form of arrays capturing information at every
t_k = k \cdot 0.001, yielding a \(24001 \times 1\) size array. This interval size proved sufficient enough to achieve
nearly continuous resolution in the models studied and had negligible impact on the accuracy. The func-
tions to be calculated were either given in closed-form, or as a fixed point equation. The closed-form
equations involved integrals that did not have analytic solutions. The trapezoidal method for approximat-
ing integrals was employed to calculate the final function as well as any intermediary functions needed.
To solve the FPE equation presented in Chapter 2, the fixed-point iteration method was employed after
a contraction algorithm was proven. Implementing this algorithm involved the double integration of a
two-parameter function. Using \(dt = 0.001\) as a scale for this approximation was infeasible as it would
require a matrix of size \(24000^2\). This approximation was done with scale \(dt = .05\). When values of
these predetermined functions were needed in the stochastic simulation at a random time \(t\), the value
was interpolated from the two elements in the array with the closest time points.
1.7.2 Discrete Event Simulation

With an appropriately high number of samples, repeated discrete event simulations over a fixed time horizon gives the average behavior of the system for all performance functions of interest. We were able to show that for a wide variety of model inputs that not only did the proposed fluid performance functions of the system closely match the simulated average behavior but that the system was controlled as desired.

Virtual customers

Unlike simulations capturing steady-state behavior, the variables of interest in our simulation are desired to be known as functions of time throughout an entire time-horizon \([0, T]\), where \(T = 24\) in this thesis. To facilitate the easy calculation of time-varying results, we implemented a technique referred to as virtual customers. These customers served two important purposes (i) to create a set sampling point of system status (i.e. queue lengths and HOL waiting times); and (ii) to capture the customer experience undergone by an arrival (i.e. experienced waiting times, and the decision to abandon).

To estimate the desired continuous processes, these virtual customers were arrived to the system at small set time intervals \(\Delta t\) throughout the entire time interval. A virtual customer was arrived and given a patience time. The virtual customer behaves in the simulation just as a regular customer would, except, when a virtual customer is selected to enter service, it instead leaves the system and the scheduling is repeated until a real customer enters service. This ensures that the virtual customer in no way impacts a real customer’s experience. When the virtual customer exits the system, the experienced waiting time is recorded as the potential waiting time and the customers decision to abandon or not was recorded. By averaging the captured data for virtual customers, across a large sample size, we were effectively able to approximate the time-dependent processes (i.e. queue lengths, HOL waiting times, abandonment probabilities, etc…). This provides a simple alternative to the computationally more intensive process of sorting arrivals into appropriate arrival bins for each sample and eliminated the potential for empty bins.

Monte Carlo estimators

Monte Carlo estimators rely on a strong law of large numbers argument whereby the summation of large number of independently simulated random variables divided by the large number is approximately the expected value of that same random variable. For instance, assume we are interested in capturing the mean behavior of a random variable \(X\) and we have the results of \(m\) independent simulation realizations
of $X$, given as $X_1, X_2, \ldots, X_m$. Then the Monte Carlo estimator $\widehat{E}[X]$ is given as the approximation of

$$\widehat{E}[X] \approx \frac{1}{m} \sum_{j=1}^{m} X_j.$$ 

Monte Carlo estimators can be extended to time-varying stochastic processes considered in this thesis because at every time $t$ the behavior of the process can be considered a single random variable to be estimated parameterized by $t$. For instance, in a simple queueing model, a stochastic process of interest is the expected potential waiting time an arrival at time $t$ would experience before being served, which we will denote $V(t)$. To approximate the process, we conduct $m$ independent discrete event simulations and capture the potential waiting time as described above with virtual customers. We will have the resulting set of realizations $V_{ij}$, where $i = 1, 2, \ldots T/\delta t$ is the time index of the realization and $j = 1, \ldots, m$ is the sample run the realization belongs to. We then have the following Monte Carlo estimator approximation for potential waiting time

$$V(t) \approx \widehat{E}[V(t)] = \frac{1}{m} \sum_{j=1}^{m} V_{i[j]},$$

(1.1)

where $[t/\delta t]$ indicates the rounding to the closest integer and represents the closest captured time index for $t$. Due to the linearity of expectation, we can also estimate functions of random variables. We take advantage of the indicator function to create probabilities involving the random processes, because the expectation of an indicator is the probability that the indicator is 1. We define an estimator for another key performance indicator, time-dependent probability of delay (PoD), $\mathbb{P}(V(t) > 0)$, as

$$\mathbb{P}(V(t) > 0) \approx \widehat{E}[1\{V([t/\delta t]) > 0\}] = \frac{1}{m} \sum_{j=1}^{m} 1\{V_{i[j]} > 0\}.$$ (1.2)

In Chapter 2 we deal extensively with the tail probability of delay $\mathbb{P}(V(t) > x)$.

**Time-varying staffing**

To simulate time-varying staffing requires careful consideration. Continuous staffing functions must first be discretized in a manner according to the discussion in §1.1. The resulting function must be searched for the time points where the function increases or decreases, indicating either the addition or removal of a server, respectively. If the waiting queue is non-empty at the time of a server addition, the appropriate next customer is then routed to the new server instantaneously. If there exists no appropriate customer to enter service, the server becomes idle. If at the time of a specified server decrease a server is idle, this server is departed from the system. Because this simulation in this proposal was done to understand the impact of total service resource on the system, the identity of the server leaving is not relevant.
If all servers are busy when a server reduction is required, a policy must be implemented on how to handle the reduction. One such policy is simply to depart one of the servers immediately along with the customer in service. However, as this work deals primarily with service system applications, the policy was implemented that the next server to finish processing a customer would exit the system. Servers are decreased in this manner until the target staffing level is met. This policy will not cause much of an impact on the ability to meet the staffing level unless service times are large.

**Implementation**

To model the random behavior of a queueing system with time-varying staffing using discrete event simulation, the core events can be separated into four main categories *customer arrival, customer service completion, customer abandonment*, and *server adjustment*. Each random event corresponds to a sequence of deterministic model functions that alter the state of the system. For the work in this proposal, we considered an additional event of *virtual customer arrival*. System state variables help keep track of the important features of the model such as number of customers in queue and service, number of servers, wait times of all customers in queue, and cumulative total number of arrivals, service completions, and abandonments.

Time-dependent simulations are often indexed by a universal time variable $t$. An event list of the next occurring time for each of the possible events is kept to determine which event is the next to handle. The computer simulation finds the next event (one with occurrence time closest to $t$) and advances $t$ until that point. Then, the deterministic actions belonging to the specific event are carried out in addition to updating the event list. For instance, when an customer enters service, the time of the next service completion must be reevaluated to include this new customer. This process continues until a stopping criteria is met. For a discrete-event simulation interested in steady-state behavior the difference in system state variables is captured to determine when the system is not changing. For transient behavior, an upper limit can be placed on $t$ to capture a time-period of interest. To conduct the Monte Carlo simulation, the discrete event simulation process is repeated until the target number of samples has been met. In this work, a large number sample size was chosen and a 99% CI was calculated to determine if a larger sample size was required.
Chapter 2

Staffing and Scheduling to Differentiate Service in Time-Varying Multiclass Service Systems

2.1 Introduction

In this chapter, we study a service-level differentiation problem for a many-server service system with (a finite number) $K$ customer classes each having its own dedicated queue and time-varying arrival rate. The problem of achieving differentiated service can be framed as concurrent determination of a staffing (i.e., number of servers) and scheduling (i.e., pairing a newly available server with a customer when there are customers from more than one class waiting) rule to satisfy a set of prescribed performance targets. In the present study, we are especially interested in satisfying the following service-level constraints:

$$
\mathbb{P} \left( V_i(t) > w_i \right) \leq \alpha_i, \quad 1 \leq i \leq K, \quad 0 < t < T,
$$

(2.1)

for class-specific delay target $w_i$ and tail-probability target $\alpha_i \in (0, 1)$, $1 \leq i \leq K$, finite time horizon $T$ (e.g., $T = 24$), where $V_i(t)$ is the delay of a class-$i$ customer arriving at time $t$. In words, the set of constraints requires that a class $i$ customer who arrives at time $t$ waits longer than $w_i$ time units with a probability no greater than $\alpha_i$. We refer to the left-hand side of (2.1) as the tail probability of delay (TPoD). Such TPoD-based quality-of-service (QoS) metrics have been widely used in service systems, such as the 80/20 rule in call centers [3] [32] and the 6-hour service level in Singapore hospitals [85].

Ideally, we would like to use the minimum possible staffing to meet those targets, in which case one expects that all the constraints in (2.1) are binding or nearly binding. Note that the minimum staffing level depends critically on the space of scheduling policies. Here, instead of solving an optimal staffing problem subject to constraints, we seek simple and effective scheduling rules that can achieve perfor-
Performance stabilization in a finite time period across all customer classes. Loosely speaking, we look for a staffing function and a scheduling policy under which

\[ P(V_i(t) > w_i) \approx \alpha_i, \quad 1 \leq i \leq K, \quad 0 < t < T. \]

From now on we refer to the above problem as the \textit{service differentiation and performance stabilization} problem.

Motivations for the present study largely arise from human-operated service systems where the system operator needs to determine how to economically plan and fairly allocate scarce service resources (e.g., number of servers) to meet the diverse needs of its customers. One notable example is the \textit{Canadian triage and acuity scale} (CTAS) guideline that classifies patients in the emergency department (ED) into five acuity levels, where each acuity level is associated with a prescribed performance target, consisting of a threshold time and the proportion of patients whose waiting time should not exceed that threshold. According to the CTAS guideline [27, 19], “CTAS level \( i \) patients need to be seen by a physician within \( w_i \) minutes \( 100\alpha_i\% \) of the time”, with

\[(w_1, w_2, w_3, w_4, w_5) = (0, 15, 30, 60, 120) \quad \text{and} \quad (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0.98, 0.95, 0.9, 0.85, 0.8).\]

In this setting, healthcare personnel represents the service resource which can be effectively staffed and scheduled to meet the CTAS targets. Similar multi-level triage policies have been widely adopted in many other EDs (not limited to those in Canada), see [31].

Other common examples of service differentiation include large customer contact centers where callers are often segmented into different classes each having a specified service deadline. Service differentiation is also important in today’s multi-media (or omni-channel) contact centers where one looks at the service level not just for the voice transactions alone, but for emails or webchat interactions. Each of these channels requires that we define what our service level is. There may have been 80% of the voice calls answered within 20 seconds, but in email that may equate to 80% of the emails responded within four hours, or 80% of the chat requests answered within 90 seconds; see [95]. In addition to customer contact centers, our modeling framework and proposed solutions may be applied to other service systems that share similar features. Examples include immigration offices in which the employees have to select cases to expedite in the face of a large backlog of immigration/permit applications as well as amusement parks where service providers have to tradeoff serving the fastpass and regular customers; see [58]. To summarize, our framework provides a useful tool to understand how scarce service resources should be allocated in the aforementioned systems where service strategies are not driven by revenue but rather less tangible aspects such as social welfare.

The multi-class multi-server queueing system considered in this chapter captures salient features of real-world service systems. First, we assume the demand function to be time varying for each class.
This assumption is primarily motivated by empirical studies showing that demand arrivals in real-world service systems typically vary strongly over time; see [38]. Second, we incorporate customer abandonments to reflect the fact that patients waiting in the ED may leave the system without being seen and callers may hang up due to prolonged waiting times. Third, we allow service times to have class-dependent service rates; this makes our model especially useful in practical settings. One example is the CTAS example where treatment times are evidently different for patients of various acuity levels.

2.1.1 Literature

Our chapter relates to two different streams of research, and we will review each in turn.

First the staffing component of our proposed solution is related to works on development of time-varying staffing functions to stabilize performance of relevant queueing systems having time-varying arrival-rate functions. The point-wise stationary approximation has been proven useful in staffing systems with shorter service times and slowly varying arrival rates; see [38] for a review. The modified-offered-load (MOL) approximation has been used to design staffing functions to control performance functions including the probability of delay (PoD), mean waiting time, and probability of abandonment (PoA). A key step of MOL is to staff according to the offered-load function of the corresponding infinite-server queue (which estimates the total service resource needed if there were no constraint on the capacity), see [47, 49, 66, 105, 103]. [29] developed a simulation-based iterative staffing algorithm (ISA) to stabilize performance of the PoD; the idea of ISA has been extended by [24] to treat TPoD. Recently, [62] developed an analytic staffing function to stabilize the TPoD and proved the corresponding asymptotic stability result. Staffing functions based on Gaussian variance approximations are studied in [78]. To the best of our knowledge, prior to our work there exists no result on joint staffing and scheduling decisions in overloaded time-varying queues.

Second, the scheduling component of our proposed solution to the service differentiation problem relates to a vast body of research on optimal scheduling control for queueing systems. Using the conventional heavy-traffic (HT) scaling, [99] showed the celebrated $c_{HT}$ rule to be asymptotically optimal; see also [72]. Similar approaches were adopted by [12, 45, 10] for critically loaded systems and by [11] for overloaded systems in the many-server setting. More recently, [52] incorporated the customer patience-time distribution into an optimal scheduling problem. Using heavy-traffic analysis, they proposed a near-optimal scheduling policies that can be implemented by customer contact centers to further improve performance metrics. In this chapter, we too use generally distributed patience time and devise control solutions that account for temporal changes in customer patience. Finally, we point out the empirical work of [27] that used patient-level data to analyze patient routing behaviors; their empirical findings suggest that the Canadian EDs apply a delay-dependent prioritization across different triage levels. The formulation of our problem is mostly related to the constraint-satisfaction approach as adopted by [40] and [42]; see also [87]. By focusing on ratio scheduling and routing policies, [42]
sought “good and simple” policies and established the state-space-collapse (SSC) associated with the HT limit showing that the ratio rules are asymptotically optimal. More recent, [90] applied ratio rules in a time-varying environment to achieve service differentiation in a critically-loaded system. It is worth noting that all of these papers assume a critical loading system (hence the delay becomes negligible as the system size increases). These results may not be applicable to systems that are operating in the overloaded regime so that customer waiting times are comparable to their service times and thus not negligible (e.g., healthcare systems).

In this chapter, we focus on the treatment of overloaded systems where the queue length and customer waiting times are no longer negligible compared to service times (In this sense, overloaded systems can be more difficult to analyze than critically loaded systems). For this reason, we do not scale down the delay targets, as was done in [42, 87, 90]. Moreover, motivated by CTAS-type service levels, we allow the probability targets $\alpha_i$ to be different across classes as opposed to identical targets as used by [42, 90]. Finally, we consider class-dependent service rates $\mu_i$ rather than a common service rate $\mu$ for all classes as was imposed by [40, 42, 87] and [52].

2.1.2 Contributions

First, we propose a new time-varying staffing rule and time-varying dynamic scheduling policy for the multiclass model, having time-varying arrivals, customer abandonment, and class-dependent service, abandonment, and arrival rates. Our dynamic scheduling rule prioritizes customers based on their elapsed delays with a time-varying class-dependent prioritization regulator. Second, under our new staffing and scheduling policies, we establish many-server heavy-traffic (MSHT) functional central limit theorem (FCLT) for various quantities of interest. By proving an SSC result for the waiting time processes, we show that the multiclass delay functions reduce to a simple one-dimensional process (called the frontier process). Because we allow service rates to be class dependent, our frontier process uniquely solves a stochastic Volterra equation, which is in sharp contrast with the existing literature wherein Ornstein-Uhlenbeck (or piecewise linear diffusion) processes often arise as the scaling limit. Third, based on the SSC and further analysis of this frontier process, we obtain the desired analytic control functions for our time-varying staffing and scheduling policies; The computation of these control functions replies on the first and second moment of the limiting frontier process for which we develop efficient algorithms. We prove that they asymptotically achieve TPoD-based service-level differentiation for all classes at customized service targets. See Figure 2.1 for an illustration of our main steps. Last, we consider important special cases to gain useful insights of our staffing and scheduling policies. We also conduct extensive simulation experiments to substantiate the effectiveness and robustness of our results.

Organization of the chapter. In §2.2, we describe the multiclass time-varying queueing model and introduce our staffing and scheduling policies. In §2.3, we present the limit theorems which establish
asymptotic service differentiation and stabilization. In §2.4 we report numerical examples. The main technical proofs are given §2.5. Finally, we make concluding remarks in §2.6.

2.2 Problem Formulation and Proposed Solutions

We describe the time-varying multiclass queueing model in §2.2.1. We introduce our time-varying staffing and dynamic scheduling rules in §2.2.2 and §2.2.3.

2.2.1 A Multiclass V model

Consider a V-model having $K \geq 2$ customer queues served by one common service pool. Customers arrive to the $i^{th}$ queue according to a non-homogeneous Poisson process (NHPP) $A_i$ with rate function $\lambda_i(\cdot)$. In what follows, we will be using $\Lambda_i(\cdot)$ to denote the corresponding cumulative arrival function, i.e., $\Lambda_i(t) \equiv \int_{t_i}^{t} \lambda_i(u) du$, if the process starts at time $t_i$. For mathematical convenience, we assume that the class-$i$ arrival process $A_i$ starts at time $-w_i$. This assumption facilitates the mathematical treatment, because the proposed scheduling policy (to be specified later) can be simply implemented at time 0. We discuss how this assumption can be relaxed in Remark 2.2.3.

We assume class-$i$ service times are independent and identically distributed (i.i.d.) random variables following an exponential distribution with class-dependent service rate $\mu_i$. Class-$i$ customers
may choose to abandon from the $i^{th}$ queue according to i.i.d. abandonment times following a general distribution, with cumulative distribution function (CDF) $F_i(x)$, complementary CDF (CCDF) $F_i^c(x) \equiv 1 - F_i(x)$, probability density function (PDF) $f_i(x)$, and hazard rate $h_{F_i}(x) \equiv f_i(x)/F_i^c(x)$. We assume that service times and patience times are mutually independent of the arrival processes. Throughout this chapter, we will assume $\lambda_i(\cdot)$ to be bounded away from zero and infinity, having a piecewise bounded first-order derivative. In addition, we assume the PDF $f_i(x) > 0$ for $x \geq 0$ so that the CCDF $F_i^c(x) > 0$ on any compact interval.

The system adopts a work-conserving policy, i.e., no customers wait in queue if there is an available server. Figure 2.2 gives a graphical illustration of a three-class system. Let $Q_i(t)$ represent the number of customers waiting in the $i^{th}$ queue. We use $E_i(t)$ and $R_i(t)$ to denote the number of customers that have entered service and that have abandoned from the $i^{th}$ queue, respectively, up to time $t$. By flow conservation

$$Q_i(t) = Q_i(0) + A_i(t) - E_i(t) - R_i(t). \quad (2.3)$$

Let $B_i(t)$ be the number of busy servers currently serving class-$i$ customers at time $t$ and $D_i(t)$ be the cumulative number of class-$i$ customers that have departed due to service completion up to time $t$. Again by flow conservation, we get

$$B_i(t) = B_i(0) + E_i(t) - D_i(t). \quad (2.4)$$

Finally, let $X_i(t)$ denote the total number of class-$i$ customers in the system at time $t$. Adding up (2.3) and (2.4) yields

$$X_i(t) = Q_i(t) + B_i(t) = X_i(0) + A_i(t) - D_i(t) - R_i(t). \quad (2.5)$$
Two waiting times. We now introduce two types of waiting-time processes that we will exploit heavily in the subsequent analysis. Let $H_i(t)$ denote the head-of-line waiting time (HWT) of the $i$th queue, i.e., the waiting time of the class-$i$ customer who has been waiting the longest (if there is any); $H_i(t) = 0$ if there is no customer waiting in the $i$th queue. Let $V_i(t)$ represent the class-$i$ potential waiting time (PWT) at time $t$, i.e., the waiting time of a potential class-$i$ customer arriving at time $t$ who has infinite patience. Based on these two waiting times, we can conveniently express the enter-service process and the queue-length process for each customer class in the following way:

$$E_i(t) = \sum_{k=1}^{A_i(t-H_i(t))} 1_{\{\gamma_{i,k} > V_i(\xi_{i,k})\}}, \quad (2.6)$$

$$Q_i(t) = \sum_{k=A_i(t-H_i(t))}^{A_i(t)} 1_{\{\xi_{i,k} + \gamma_{i,k} > t\}}, \quad (2.7)$$

where $1_A$ denotes the indicator function of event (set) $A$, the random variables $-w_i \leq \xi_{i,1} < \xi_{i,2} < \cdots$ denote the successive arrival times of class-$i$ customers, and $\gamma_{i,1}, \gamma_{i,2}, \ldots$ denote the i.i.d. patience times with CDF $F_i$. As will become clear in the subsequent analysis, these representations are useful in deriving the limiting FCLT results. To complete the model, it remains to specify (i) the staffing level for the service pool (which plans the overall service capacity for all customer classes), and (ii) the scheduling policy used to pair a newly available server with a waiting customer from one of $K$ classes (which determines how to dynamically allocate the overall service capacity to serve each customer class).

2.2.2 A Time-Varying Square-Root Staffing Rule

Following Step 1 in Figure 2.1, we now introduce a time-varying square-root staffing (TV-SRS) rule, which consists of two terms: (i) the nominal staffing level (first-order term) and (ii) the safety staffing level (second-order term).

First-order nominal staffing term. To set the nominal staffing level, we adopt the offered load analysis which estimates the required service capacity by estimating how much capacity would be used if there were no limit on its availability. For example, consider a single-class $Mt/GI/s_t + GI$ model having Poisson arrivals rate $\lambda(t)$, independent and i.i.d. service times with a general distribution $G$ (the first $GI$), and i.i.d. customer abandonment following a general distribution $F$ (the $+GI$). Although the $Mt/GI/s_t + GI$ model is complicated, the corresponding $Mt/GI/\infty$ infinite-server model remains remarkably tractable, where the number of customers (or busy servers) follows a Poisson distribution.
with mean

\[ m_\infty(t) \equiv \mathbb{E}[X_\infty(t)] = \int_0^t \lambda(u)G^c(t-u)du. \] (2.8)

If the objective is to stabilize the expected delay at any given point in time at a target \( w \), one will need to set the staffing levels to a modified version of (2.8), namely,

\[ m_{\text{DIS}}(t) \equiv \int_0^t F^c(w)\lambda(u-w)G^c(t-u)du, \] (2.9)

where we have used DIS to denote the “delayed-infinite-server approximation”, as in [66]. The effective arrival rate can be justified by the fact that, if every arrival who does not elect to abandon waits \( w \) time units, then a fraction \( F(w) \) of arrivals will abandon the queue before entering service. In other words, one can think of \( m_{\text{DIS}}(t) \) as the \emph{mean number of busy servers needed to serve all customers who are willing to wait for \( w \) time units}.

For our multiclass V model with class-dependent delay \( w_i \), we follow the above offered-load analysis by setting the nominal staffing level as

\[ m(t) \equiv \sum_{i=1}^K m_i(t), \quad \text{where} \quad m_i(t) \equiv \int_0^t F^c_i(w_i)\lambda_i(u-w_i)\exp(-\mu_i(t-u))du, \] (2.10)

where each term in the sum of (2.10) is obtained by replacing \((F, w, G, \lambda)\) in (2.9) with the class-dependent primitives \((F_i, w_i, \exp(\mu_i), \lambda_i)\).

**Second-order safety staffing term.** Unfortunately, \( m(t) \) is not effective for stabilizing class-dependent TPoDs, because \( m(t) \) does not include the class-dependent probability targets \( \alpha_i \). Our strategy is to refine the staffing level by adding a second-order safety staffing term that is a function driven by the class-dependent probability targets \( \alpha_i \). Let \( \lambda(t) \equiv \sum_{i=1}^K \lambda_i(t) \) be the aggregate demand function, and \( \bar{\lambda} \equiv T^{-1} \int_0^T \lambda(t)dt \) be the average arrival rate over \([0, T]\). We envision a staffing function consisting of two pieces, namely,

\[ s(t) = \left\lceil m(t) + \sqrt{\bar{\lambda}}c(t) \right\rceil, \] (2.11)

where \( \lceil x \rceil \) is smallest integer that is greater than or equal to \( x \), and \( c(t) \equiv c(t, \alpha_1, \ldots, \alpha_K) \) is a time-varying and \((\alpha_1, \ldots, \alpha_K)\)-dependent piecewise continuous control function, which will be determined later. We refer to such a staffing formula (2.11) as TV-SRS.

**Remark 2.2.1 (Role of the Safety Staffing Functions)** Note that the first-order nominal term \( m(t) \) in (2.11) lives on the order of \( \bar{\lambda} \), while the second-term term lives on the order \( \sqrt{\bar{\lambda}} \). Given that the offered load \( m(t) \) depends on delay target \( w_i \), arrival rate \( \lambda_i(t) \), service rate \( \mu_i \) and patience-time
distribution $F_i$, the remaining flexibility in the staffing formula depends entirely on the single control function $c$, which will be determined to satisfy the performance targets, as specified by (2.1). Hence, the overall staffing level $s$ depends on probability targets $(\alpha, \ldots, \alpha_K)$ only through $c$.

2.2.3 A Time-Varying Dynamic Prioritization Scheduling Rule

Following Step 1 in Figure 2.1, we next introduce a delay-based dynamic scheduling rule which is both time dependent and state dependent. To implement such a scheduling policy, we track the elapsed waiting time of all waiting customers. Because customers are served under FCFS within each class, it suffices to track the HWTs, namely, $(H_1(t), \ldots, H_K(t))$.

We route the next class-$i^*$ head-of-line (HoL) customer (if any) into service, with $i^*$ satisfying

$$i^* \in_{1 \leq i \leq K} \left\{ \frac{H_i(t)}{w_i} \text{ normalized HWT} + \frac{1}{\sqrt{\lambda}} \kappa_i(t) \right\},$$

(2.12)

where the first term $H_i(t)/w_i$ is the HWT scaled by the delay target, and $\kappa_i(t) \equiv \kappa_i(t, \alpha_i)$, referred to as the second-order class-$i$ prioritization regulator, is a time-varying and $\alpha_i$-dependent piecewise continuous control function to be specified later. We refer to such a scheduling rule as the time-varying dynamic prioritization scheduling (TV-DPS) policy. Furthermore, we define what we call the frontier process as

$$H(t) \equiv \frac{H_{i^*}(t)}{w_{i^*}} + \frac{1}{\sqrt{\lambda}} \kappa_{i^*}(t).$$

(2.13)

Remark 2.2.2 (Understanding TV-DPS) The first-order term $H_i(t)/w_i$ is designed to guarantee that the class-$i$ delay is close to its target $w_i$ (it is controlling the relative delay imbalance $(H_i(t) - w_i)/w_i$, rather than the absolute delay imbalance). The idea of exploiting the head-of-line delay information dates back to [57]; see also [60] for a non-linear extension. The second-order term $(1/\sqrt{\lambda})\kappa_i(\cdot)$ helps accomplish the class-dependent probability target $\alpha_i$. Intuitively, such a control function $\kappa_i$ should satisfy the following properties:

(i) Monotonicity. For fixed time $t$, $\kappa_i(t)$ should be a decreasing function of $\alpha_i$, because a bigger value of $\alpha_i$ means a lower service quality, which yields a lower prioritization level for class $i$;

(ii) Sign. For a class $i$ with probability target $\alpha_i > 0$ ($\alpha_i \leq 0.5$), the fine-tuning prioritization regulator $\kappa_i$ should satisfy $\kappa_i(t) < 0$ ($\kappa_i(t) \geq 0$) (Benchmarking with the case $\alpha_i = 0.5$, $\kappa_i$ should base on the value of $\alpha_i$ to adjust the priority levels by adding a positive or negative weight to $H_i(t)/w_i$). See numerical examples in §2.4 for more discussions of the structure of $\kappa_i$.

Our TV-DPS rule is both time dependent (accounting for time variability in the arrival processes) and state dependent (dynamically capturing the system’s stochasticity). To the best of our knowledge,
this is a feature unique to the present study and absent from previous research. Moreover, our proposed scheduling policy is in alignment with the current practice of Canadian EDs where patients are routed not only by triage level (static) priorities but also by their actual (dynamic) wait time, as documented by [27]. This makes this rule especially appealing as the intrinsic fairness of the TV-DPS policy helps achieve ethical expectations set forth by the CTAS guideline. Furthermore, when $w_i = w$ and $\alpha_i = \alpha$ for all $1 \leq i \leq K$, TV-DPS degenerates to the global FCFS scheduling policy.

**Remark 2.2.3 (Relaxation of the assumption of different arrival times)** We assumed that each class-$i$ arrival process begins at a different (negative) time $-w_i$, so that by time 0 (at which we begin to serve all customers following TV-SRS and TV-DPS) we already have enough candidate customers. More important, each class-$i$ HoL customer is “old” enough (meaning they have reached the specific class-$i$ delay target $w_i$). This provide a clean condition for our mathematical treatment.

We now briefly discuss the situation where customers of each class start to arrive at time zero. Suppose there are three classes with delay targets $w_1$, $w_2$, and $w_3$, respectively. Without loss of generality, we assume $w_1 < w_2 < w_3$. Then a modified version of the TV-DPS rule proceeds as follows. Over the period $[0, w_1)$, we do not serve any customers. During $[w_1, w_2)$, we act as if there is one customer class, namely, class 1. During $[w_2, w_3)$, we pretend that there are only two classes, namely class 1 and 2 (i.e., choose to serve the first two classes only), and apply the rule (2.13) for $K = 2$. At time $w_3$ and beyond, the TV-DPS rule is implemented in the usual way for all classes.

In the next section, we will first establish an MSHT FCLT result under our TV-SRS and TV-DPS rules with unknown control parameters $c$ and $\kappa_i$ (Step 2 in Figure 2.1); using the FCLT limit, we will next obtain the exact formulas of $c$ and $\kappa_i$ so that the TPoD-based service-level constraints are asymptotically satisfied as the scale increases (Step 3 in Figure 2.1).

### 2.3 Asymptotic Service-Level Differentiation

In this section, we present our main results. §2.3.1 gives the asymptotic framework and states the many-server FCLT and FWLLN results for the multiclass V model operating under the TV-SRS and TV-DPS policies introduced in §§2.2.2–2.2.3. In §2.3.2 we utilize the FCLT results to obtain the desired control factors $\kappa_i$ and $c$ and show that they asymptotically achieve TPoD-based service-level differentiation and performance stabilization. §2.3.3 provides a more detailed discussion of the important case of class-independent service rate. All proofs are given in §2.5.

#### 2.3.1 Many-Server FCLT Limits under TV-SRS and TV-DPS

To establish an FCLT limit, we consider an asymptotic framework in which the system scale (here the average arrival $\bar{\lambda}$) grows to infinity. Following the convention in the literature, we will use $n$ in place
of $\bar{\lambda}$ as our scaling parameter. This gives rise to a sequence of $K$-class V models indexed by $n$. Let $A_i^n(t)$ be the class-$i$ NHPP arrival process in the $n^{th}$ model, having a rate function $n\lambda_i(\cdot)$ where, by slight abuse of notation, we used $\lambda_i(t)$ to denote the baseline arrival rate at time $t$. Our TV-SRS function satisfies
\begin{equation}
 s^n(t) = \lceil nm(t) + \sqrt{n}c(t) \rceil, \tag{2.14}
\end{equation}
where $m$ and $c$ are the offered-load function in (2.10) and safety staffing term (yet to be determined).

Let $H_i^n$ and $V_i^n$ be the class-$i$ HWT and PWT in the $n^{th}$ model. Our TV-DPS satisfies
\begin{equation}
 i^* \in \{1 \leq i \leq K \} \left\{ \frac{H_i^n(t)}{w_i} + \frac{1}{\sqrt{n}} \kappa_i(t) \right\}, \tag{2.15}
\end{equation}
where $\kappa_i$ is a control function yet to be determined.

For $1 \leq i \leq K$, let
\begin{equation}
 \Lambda_i(t) \equiv \int_{-w_i}^{t} \lambda_i(u)du, \quad \tilde{A}_i^n(t) \equiv n^{-1}A_i^n(t) \quad \text{and} \quad \hat{A}_i^n(t) \equiv n^{-1/2} \left( A_i^n(t) - n \Lambda_i(t) \right). \tag{2.16}
\end{equation}
The sequence of processes $\tilde{A}_i^n$ and $\hat{A}_i^n$ satisfy a FWLLN and FCLT, namely,
\begin{equation}
 (\tilde{A}_i^n(\cdot), \hat{A}_i^n(\cdot)) \Rightarrow (\Lambda_i(\cdot), \hat{A}_i(\cdot)) \quad \text{in} \quad \mathcal{D}^2 \quad \text{as} \quad n \to \infty, \tag{2.17}
\end{equation}
for $\hat{A}_i(\cdot) \equiv W_{\lambda_i} \circ \Lambda_i(\cdot)$, where $x \circ y(t) \equiv x(y(t))$, $W_{\lambda_i}$ being a standard Brownian motion, and $\mathcal{D} \equiv \mathcal{D}(\mathbb{R}_+, \mathbb{R})$ is the space of right-continuous $\mathbb{R}$-valued functions on $\mathbb{R}_+$ with lefthand limit, which is endowed with the Skorokhod $J_1$-topology, and $\Rightarrow$ means convergence in distribution (weak convergence).

**Remark 2.3.1 (More general $G_t$ arrivals)** Our main results below can be easily extended to more general $G_t$ arrival processes (which are not necessarily NHPPs), as long as their CLT-scaled versions satisfy the FCLT
\begin{equation}
 \hat{A}_i^n(\cdot) \Rightarrow c_{\lambda_i}W_{\lambda_i} \circ \Lambda_i(\cdot) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty,
\end{equation}
for some $c_{\lambda_i} > 0$. These types of $G_t$ arrival processes can be used to model over-dispersed and under-dispersed arrival processes (i.e., when the variance-to-mean ratio of the number of arrivals is not close to 1), see [35, 47, 61, 62] for construction and analysis of such $G_t$ arrival processes. In this case, our FCLT limits in Theorem 2.3.1 can be easily adjusted by simply multiplying $W_{\lambda_i}$ by the constant $c_{\lambda_i}$. For NHPPs, $c_{\lambda_i} = 1$.

Following the notations in §2.2.1, we use $Q_i^n(t)$ and $B_i^n(t)$ to denote the number of class-$i$ customers in queue and in service at time $t$, respectively in the $n^{th}$ V model. Their sum, denoted by $X_i^n(t)$,
represents the total number of class-\(i\) customers in system at time \(t\). We now define their corresponding CLT-scaled versions

\[
\begin{align*}
\tilde{B}_i^n(t) &\equiv n^{-1/2} \left( B_i^n(t) - nm_i(t) \right), \\
\tilde{Q}_i^n(t) &\equiv n^{-1/2} \left( Q_i^n(t) - nq_i(t) \right), \\
\tilde{X}_i^n(t) &\equiv n^{-1/2} \left( X_i^n(t) - nx_i(t) \right),
\end{align*}
\]

for \(m_i\) given by (2.10) and \(q_i(t) \equiv \int_{t-w_i}^{t} F_i^c(t-u) \lambda_i(u)du\) respectively and \(x_i \equiv m_i + q_i\). We let

\[
\tilde{H}_i^n(t) \equiv n^{1/2} (H_i^n(t) - w_i) \quad \text{and} \quad \tilde{V}_i^n(t) \equiv n^{1/2} (V_i^n(t) - w_i)
\]

be the CLT-scaled HWT and PWT processes, respectively. Finally, define the CLT-scaled frontier process

\[
\tilde{H}^n(t) \equiv n^{1/2} (H^n(t) - 1).
\]

**Theorem 2.3.1 (MSHT FCLT limits under TV-SRS and TV-DPS)** Suppose the system operates under TV-SRS in (2.14) and TV-DPS in (2.15). Then there is a joint convergence for the CLT-scaled processes:

\[
\left( \tilde{H}^n, \tilde{B}^n_1, \ldots, \tilde{B}^n_K, \tilde{H}^n_1, \ldots, \tilde{H}^n_K, \tilde{V}^n_1, \ldots, \tilde{V}^n_K, \tilde{X}^n_1, \ldots, \tilde{X}^n_K, \tilde{Q}^n_1, \ldots, \tilde{Q}^n_K \right)
\Rightarrow \left( \tilde{H}, \tilde{B}_1, \ldots, \tilde{B}_K, \tilde{H}_1, \ldots, \tilde{H}_K, \tilde{V}_1, \ldots, \tilde{V}_K, \tilde{X}_1, \ldots, \tilde{X}_K, \tilde{Q}_1, \ldots, \tilde{Q}_K \right) \quad \text{in} \quad D^{k+1},
\]

as \(n \to \infty\), where the FCLT limits on the right-hand side are well-defined stochastic processes.

(i) The limiting processes \((\tilde{H}, \tilde{B}_1, \ldots, \tilde{B}_K)\) jointly satisfy the following set of \(K\) Ornstein-Uhlenbeck (OU) type stochastic integral equations

\[
\tilde{B}_i(t) + \nu_i(t)\tilde{H}(t) = - \int_0^t \mu_i \tilde{B}_i(u)du - \int_0^t \psi_i(u)\tilde{H}(u)du + \int_0^t \psi_i(u)\kappa_i(u)du + \eta_i(t)\kappa_i(t) + G_i(t) \quad \text{for} \quad i = 1, \ldots, K, \quad \text{and} \quad \sum_{i=1}^K \tilde{B}_i(t) = c(t),
\]

where \(\eta_i(t) \equiv w_i \lambda_i(t-w_i) F_i^c(w_i), \psi_i(t) \equiv w_i \lambda_i(t-w_i) f_i(w_i), \)

\[
G_i(t) \equiv \tilde{E}_{i,1}(t) + \tilde{E}_{i,2}(t) - \tilde{D}_i(t), \quad \tilde{E}_{i,1}(t) \equiv F_i^c(w_i) \int_{-w_i}^t \sqrt{\lambda_i(u)}dW_{\lambda_i}(u), \quad \tilde{E}_{i,2}(t) \equiv \int_{-w_i}^t \sqrt{\lambda_i(u)}dW_{\kappa_i}(u), \quad \tilde{D}_i(t) \equiv \int_0^t \sqrt{\mu_i m_i(u)}dW_{\mu_i}(u),
\]

and \(W_{\lambda_i}, W_{\kappa_i}, W_{\mu_i}\) are independent standard Brownian motions.
The FCLT limits for all HWT and PWT processes are deterministic functionals of a one-dimensional process $\hat{H}$, namely,

$$\hat{H}_i(t) \equiv w_i(\hat{H}(t) - \kappa_i(t)), \quad \text{and} \quad \hat{V}_i(t) = w_i(\hat{H}(t + w_i) - \kappa_i(t + w_i)); \quad (2.22)$$

The FCLT limit for each queue-length process is the sum of three terms, namely, $\hat{Q}_i(t) \equiv \hat{Q}_{i,1}(t) + \hat{Q}_{i,2}(t) + \hat{Q}_{i,3}(t)$, where

$$\hat{Q}_{i,1}(t) \equiv \int_{t-w_i}^{t} F_i^c(t-u) \sqrt{\lambda_i(u)} dW_{\lambda_i}(u),$$

$$\hat{Q}_{i,2}(t) \equiv \int_{t-w_i}^{t} \sqrt{F_i^c(t-u)F_i(t-u)\lambda_i(u)} dW_{\theta_i}(s),$$

$$\hat{Q}_{i,3}(t) \equiv \lambda_i(t-w_i)F_i^c(w_i)\hat{H}_i(t).$$

Finally, the FCLT limits for number in system is given by $\hat{X}_i(t) = \hat{B}_i(t) + \hat{Q}_i(t)$.

Remark 2.3.2 (SSC and Separation of Variability) Theorem 2.3.1 provides the FCLT limits for waiting times and queue lengths under TV-SRS and TV-DPS with the second-order terms $c$ and $\kappa_i$ yet to be determined. Such FCLT results will be used later to achieve asymptotic performance differentiation and stabilization. Part (ii) of Theorem 2.3.1 gives a nice SSC result: The diffusion limits $(\hat{H}, \hat{B}_1, \ldots, \hat{B}_K)$ satisfy the $(K+1)$-dimensional stochastic differential equation (SDE), and according to (A.2), both limiting HWT and PWT processes are deterministic functionals of the one-dimensional limiting frontier process $\hat{H}$. The intuition behind the SSC is that all these normalized HWTs (plus the second-order prioritization regulator) in (2.13) do not differ much from each other under the TV-DPS policy. In fact, the difference between any two classes is of order $O(1/n)$. In addition, there are $3K$ independent Brownian motions $W_{\lambda_i}, W_{\theta_i}, W_{\mu_i}$, stemming from the independent random sources (arrival, abandonment and service) of all $K$ customer classes. We will see later in Proposition 1, these sources of randomness jointly contribute to the variability of the one-dimensional process $\hat{H}$.

We next provide an FWLLN result for the V model operating under the TV-SRS and TV-DPS rule. For that purpose, we define the LLN-scaled processes as follows

$$\bar{B}_i^n(t) \equiv n^{-1} B_i^n(t), \quad \bar{Q}_i^n(t) \equiv n^{-1} Q_i^n(t), \quad \bar{X}_i^n(t) \equiv n^{-1} X_i^n(t) \quad \text{for} \quad 1 \leq i \leq K. \quad (2.23)$$

The next result is a direct consequence of Theorem 2.3.1.

Corollary 2.3.1 (FWLLN) Suppose that the system operates under TV-SRS in (2.14) and TV-DPS in
There exists a time-varying number of customers (of all types) waiting to be processed at any given point in time, where those who will later abandon are included.

Asymptotically, a customer enters service at $t$ only when he arrived at $t - w_i$, and that the fraction of customers who do not abandon during $w_i$ is $F_i^c(w_i)$. Hence, according to Little’s law, $\eta(t)$ can be interpreted as the time-varying number of customers (of all types) waiting to be processed at $t$, excluding those who will later abandon.

Below we provide a proof sketch of the theorem. The details are given in §2.5.

**Proof sketch of Theorem 2.3.1.**

Step 1: We first show that each component within the curly bracket in (2.15) is at most $O(1/\sqrt{n})$ away from the frontier process, that is, $H_i^n(t)/w_i + n^{-1/2}\kappa_i(t) = H^n(t) + O(1/n)$ (or $\hat{H}_i^n(t) = w_i(H^n(t) - \kappa_i(t)) + O(1/\sqrt{n})$). This is essentially a SSC result and follows from a key observation that, at any given point in time, the number of total departures required for a HoL customer to enter service under the TV-DPS policy is of order $O(n)$.

Step 2: We then use (2.6) to obtain a simple relation between $\hat{H}_i^n$ and $\hat{B}_i^n$. Based on the fact that the difference between $\hat{H}_i^n(t)$ and $w_i(H^n(t) - \kappa_i(t))$ can be made arbitrarily small for $n$ large enough, we are able to establish a set of $K$ differential equations and one linear equation jointly satisfied by $(\hat{B}_1^n, \ldots, \hat{B}_K^n, \hat{H}^n)$. This allows us to apply the Gronwall’s inequality to establish the stochastic boundedness of the sequence $\{(\hat{B}_1^n, \ldots, \hat{B}_K^n, \hat{H}^n); n \in \mathbb{N}\}$, which in turn enables us to deduce the desired FWLLN results.

Step 3: An application of the continuous mapping theorem with the established FWLLN allows us to establish the Brownian limits given in (2.21) for the corresponding CLT-scaled processes. Applying the continuous mapping theorem again with these Brownian limits yields the joint convergence of $\{(\hat{B}_1^n, \ldots, \hat{B}_K^n, \hat{H}^n)\}$. Next, the FCLT for the HWT and PWT processes follows by converging-together lemma with the established FCLT for the frontier process. Step 4: Finally, the FCLT for the queue-length processes follows by first exploiting the relation between $Q_i^n$ and $H_i^n$ and then applying the continuous mapping theorem.

We next take a closer look at the dynamics of the limit frontier process $\hat{H}$. Define

$$\eta(t) \equiv \sum_{i=1}^K \eta_i(t) = \sum_{i=1}^K w_i \lambda_i(t - w_i) F_i^c(w_i).$$

Asymptotically, a customer enters service at $t$ only when he arrived at $t - w_i$, and that the fraction of customers who do not abandon during $w_i$ is $F_i^c(w_i)$. Hence, according to Little’s law, $\eta(t)$ can be interpreted as the time-varying number of customers (of all types) waiting to be processed at $t$, excluding those who will later abandon.

Note that each equation in Theorem 2.3.1 allows us to write $\hat{B}_i$ as a function of $\hat{H}$. Plugging them into the equation $\sum_{i=1}^K \hat{B}_i(t) = c(t)$ plus some algebraic simplifications yields the result below.
Proposition 1 (Distribution of the frontier process $\hat{H}$) The process $\hat{H}$ uniquely solves the following stochastic Volterra equation (SVE)

$$\hat{H}(t) = \int_0^t L(t, s)\hat{H}(s)ds + \int_0^t J(t, s)dW(s) + K(t), \quad (2.26)$$

where

$$L(t, s) \equiv \sum_{i=1}^{K} \frac{\eta_i(s)e^{\mu_i(s-t)} (\mu_i - h_{F_i}(w_i))}{\eta(t)},$$

$$J(t, s) \equiv \sqrt{\sum_{i=1}^{K} e^{2\mu_i(s-t)} (F_i^c(w_i)\lambda_i(s - w_i) + \mu_i m_i(s))},$$

$$K(t) \equiv \sum_{i=1}^{K} \left( \frac{\eta_i(t)\kappa_i(t) - \int_0^t \eta_i(s)e^{\mu_i(s-t)} (\mu_i - h_{F_i}(w_i)) \kappa_i(s)ds}{\eta(t)} - c(t) \right). \quad (2.27)$$

$W$ is a standard Brownian motion. In addition, $\hat{H}$ is a Gaussian process with

(i) mean $M_{\hat{H}}(t) \equiv \mathbb{E}[\hat{H}(t)], \quad 0 \leq t \leq T,$ uniquely solving the fixed-point equation (FPE)

$$M_{\hat{H}} = \Gamma(M_{\hat{H}}), \quad \text{where} \quad \Gamma(M_{\hat{H}})(t) \equiv \int_0^t L(t, s)M_{\hat{H}}(s)ds + K(t), \quad (2.28)$$

(ii) covariance $C_{\hat{H}}(t, s) \equiv \text{Cov}(\hat{H}(t), \hat{H}(s)), \quad 0 \leq s, t \leq T,$ uniquely solving the FPE

$$C_{\hat{H}} = \Theta(C_{\hat{H}}),$$

where the operator $\Theta$ is defined as

$$\Theta(C_{\hat{H}})(t, s) \equiv - \int_0^t \int_0^s L(t, u)L(s, v)C_{\hat{H}}(u, v)dvdudu$$

$$+ \int_0^t L(t, u)C_{\hat{H}}(u, s)du + \int_0^s L(s, v)C_{\hat{H}}(t, v)dv + \int_0^{s/t} J(t, u)J(s, u)du. \quad (2.29)$$

The FCLT for $\hat{H}$ satisfies a SVE rather than an ordinary SDE which is more commonly seen in the literature. This is solely because the service rates are assumed to be class-dependent. We summarize our key findings regarding the SVE in Remark 2.3.3.

Remark 2.3.3 (A closer look at the SVE (2.26))

(i) Analytic solutions in special cases. Such an SVE (2.26) in general has no analytic solution, except for some special cases. For example, if $\mu_i = h_{F_i}(w_i)$ for all $1 \leq i \leq K$ so that the drift term
Another important case is when \( L(t, s) \) and \( J(t, s) \) are separable functions in \( t \) and \( s \), which is the case when service rates are class independent (see §2.3.3 for discussions of this important special case).

(ii) **Variability.** The SVE is driven by the Brownian motion \( W \), which rises from aggregating all \( 3K \) independent Brownian motions \( W_{\lambda_i}, W_{\theta_i}, W_{\mu_i} \), \( 1 \leq i \leq K \) in (2.21); see the proof of Proposition 1 in §2.5 for details. Indeed, the stochastic variability of the frontier waiting time process is collectively determined by the randomness in the arrivals, service times and abandonment times.

(iii) **Dependence on control functions.** The terms \( L \) and \( J \) are functions of model inputs \( (\lambda_i, F_i, \mu_i, w_i) \) only, thus independent of the control functions \( \kappa_i \) and \( c \), which only appears in \( K \). Hence, varying \( \kappa_i \) and \( c \) will affect the mean of \( \hat{H} \), but not its variance. This is a crucial observation, because as will become clear later in §2.3.2, (i) computing the variance of \( \hat{H} \) (which is uncontrollable via \( c \) and \( \kappa_i \)) and (ii) appropriately shifting the mean of \( \hat{H} \) (by adjusting our control functions) are critical in achieving desired class-dependent service levels.

(iv) **Algorithms.** We prove Proposition 1 by showing that the operators \( \Gamma \) and \( \Theta \) are both contractions in appropriate functional spaces, see §2.5. In addition, our proof naturally leads to effective numerical algorithms for computing \( M_H \) and \( C_H \) (in fact, our algorithms converge geometrically fast). See Remark 2.5.2 for detailed discussions. It is obvious that \( M_H(t) = 0 \) if \( K(t) = 0 \) (because a zero function now solves the FPE (A.5)). This will indeed be the case considered later in §2.3.2.

### 2.3.2 Asymptotic Service Differentiation and Stabilization

Given the SSC achieved by TV-SRS and TV-DPS, we now focus on investigating the one-dimensional process \( \hat{H} \). When \( n \) is large we hope to satisfy

\[
\alpha_i \equiv P(\hat{V}_i^n(t) > w_i) = P(\hat{V}_i^n(t) > 0) \approx P(\hat{V}_i(t) > 0) = P(\hat{H}(t + w_i) - \kappa_i(t + w_i) > 0) \\
= P \left( \mathcal{N} \left( M_H(t + w_i), \sigma_H^2(t + w_i) \right) > \kappa_i(t + w_i) \right) = P \left( \mathcal{N}(0, 1) > \frac{\kappa_i(t + w_i) - M_H(t + w_i)}{\sigma_H(t + w_i)} \right) \tag{2.30}
\]

for all \( t \geq -w_i \), where \( \mathcal{N}(\mu, \sigma^2) \) is a normal random variable with mean \( \mu \) and variance \( \sigma^2 \), \( \sigma_H(t) = \sqrt{\text{Var}(\hat{H}(t))} = \sqrt{C_H(t, t)} \) is the standard deviation of \( \hat{H}(t) \) at \( t \). Equation (2.30) further simplifies to

\[
P \left( \mathcal{N}(0, 1) > \frac{\kappa_i(t) - M_H(t)}{\sigma_H(t)} \right) \approx \alpha_i, \quad t \geq 0, \tag{2.31}
\]
However, to be rigorous, we should say that the prioritization regulators
See §2.5 for the proof for Proposition 2. The safety staffing term
\[ z_\alpha = \text{the } \alpha\text{-quantile of a standard Gaussian random variable}, \text{ that is, } z_\alpha \text{ satisfies } \mathbb{P}(\mathcal{N}(0, 1) \leq \alpha) = z_\alpha. \]

One obvious solution to (2.32) is that, for any \( \kappa_i \), we can choose \( c \) appropriately so that \( K(t) \) in (A.4) is set to 0, so that \( M_\widetilde{H}(t) = 0 \) for all \( t \) (note that FPE (A.5) now has a unique solution \( M_\widetilde{H}(t) = 0 \) when \( K(t) = 0 \), and this leads to the control formulas below in (A.9)–(A.10). We next show that these control functions are indeed the unique solutions to (2.32).

**Proposition 2 (Asymptotically unique control functions)** The condition (2.32) is satisfied if and only if
\[
\begin{align*}
\eta_i(t)\kappa_i(t) - \int_0^t \eta_i(s)e^{\mu_i(s-t)}(\mu_i - h_{F_i}(w_i))\kappa_i(s)ds \\
\kappa_i(t) = z_{1-\alpha_i}\sigma_\widetilde{H}(t), & \quad 1 \leq i \leq K.
\end{align*}
\]
\[ (\text{2.33}) \text{ and (2.34)} \]

See §2.5 for the proof for Proposition 2. The safety staffing term \( c \) is indeed uniquely given by (A.9). However, to be rigorous, we should say that the prioritization regulators \( \kappa_1, \ldots, \kappa_K \) are unique up to adding any common function \( \Delta \), that is, applying any \( \hat{\kappa}_i(t) = \kappa_i(t) + \Delta(t) \) for \( 1 \leq i \leq K \) which will not make a difference in our TV-DPS rule.

**Remark 2.3.4 (Structure of the control functions)** The main idea here is that we choose appropriate control functions \( c \) and \( \kappa_i \) to tilt the mean of the error term \( \hat{\theta}_i^n(t) \) (rather than the mean of \( V_i^n(t) \)), so that asymptotically the probability mass of \( \{\hat{\theta}_i^n(t) > 0\} \) (or \( \{V_i^n(t) > w_i\} \)) can be set to desired \( \alpha_i \) at all time \( t \). We observe from (A.9) that the second-order safety staffing term \( c \) depends on \( \alpha_i \) through the second-order prioritization regulator \( \kappa_i \), and \( \kappa_i \) depends on \( \alpha_i \) through \( z_{\alpha_i} \). Consistent with Remark 2.2.2, \( \kappa_i \) is decreasing in \( \alpha_i \), and its sign depends on how \( \alpha_i \) compares with 0.5, that is, \( \kappa_i(t) > 0 \) (\( \kappa_i(t) < 0 \)) if \( \alpha_i < 0.5 \) (\( \alpha_i > 0.5 \)). When the probability target \( \alpha_i = 0.5 \) for all \( 1 \leq i \leq K \), we have \( c(t) = \kappa_i(t) = 0 \) so that we lose the second-order terms in both TV-SRS and TV-DPS formulas. Another interesting observation is that a bigger system variability leads to more contrasting prioritization standards. To elaborate, consider the case \( \alpha_1 < 0.5 < \alpha_2 \) so that \( z_{1-\alpha_1} > 0 > z_{1-\alpha_2} \) and \( \kappa_1(t) > 0 > \kappa_2(t) \), the difference of the two prioritization regulators \( \kappa_1(t) - \kappa_2(t) > 0 \) is increasing in \( \sigma_\widetilde{H}(t) \), which characterizes the system’s overall stochastic variability (recall from Remark 2.3.3 that the variability of \( \widetilde{H} \) captures the randomness of all events, including arrivals, service times and abandonment times). This implies that in a more random environment, we rely less (more) on the state-dependent portion (deterministic control regulator) of TV-DPS to inform the scheduling decision;
as the system environment becomes more volatile, information of the system state becomes less useful. Finally, we emphasize that $w_i (\alpha_i)$ is the first-order (second-order) QoS target, because a slight change in $w_i (\alpha_i)$ affects the first-order (second-order) term in both the TV-SRS and TV-DPS formulas.

The next theorem establishes the asymptotic effectiveness of our methods.

**Theorem 2.3.2 (Asymptotic service differentiation and performance stabilization)** Under TV-SRS (2.14) and TV-DPS (2.15) with $c_i(\cdot)$ and $\kappa_i(\cdot)$ specified in (A.9) and (A.10), we have the following asymptotic stability results:

(i) Mean PWT and mean HWT are both asymptotically stabilized for all classes:

$$E[V^n_i(t)] \to w_i \quad \text{and} \quad E[H^n_i(t)] \to w_i \quad \text{as} \quad n \to \infty, \quad \text{for} \quad 1 \leq i \leq K, \quad 0 < t \leq T. \quad (2.35)$$

(ii) TPoDs for PWT and HWT are both asymptotically stabilized for all classes:

$$P(V^n_i(t) > w_i) \to \alpha_i \quad \text{and} \quad P(H^n_i(t) > w_i) \to \alpha_i \quad \text{as} \quad n \to \infty, \quad \text{for} \quad 1 \leq i \leq K, \quad 0 < t \leq T. \quad (2.36)$$

**Remark 2.3.5 (Differentiation of the mean PWT)** If the goal is to asymptotically differentiate and stabilize the mean PWT $E[V^n_i(t)]$ rather than the TPoD, then we are free to choose any arbitrary $\alpha_i$ in (A.9) and (A.10), because the second-order terms of TV-SRS and TV-DPS play a negligible role as the scale $n$ increases. However, for a finite $n$, numerical examples show that it is helpful to set $\alpha_i = 0$ for all $1 \leq i \leq K$, so that the probability mass $\{V^n_i(t) > w_i\} \approx 0$ and the symmetric structure of the nearly Gaussian distribution “guarantees” a balanced mean at $w_i$. See §2.4.5 more discussions and numerical examples.

**2.3.3 The Case of Class-Independent Service Rate**

It is well known that the case of class-dependent service rate can be more complex, see [52] for example. In this subsection, we assume that service rates are class independent, that is, $\mu_i = \mu$ for all $1 \leq i \leq K$. Under this assumption, we show that the results are simplified significantly; indeed, the functions $L$ and $K$ are now separable in $t$ and $s$ so that the SVE in (2.26) degenerates to a much more tractable Ornstein-Uhlenbeck (OU) process with time-varying drift and volatility. We summarize our results below.

**Corollary 2.3.2 (Frontier process $\hat{H}$ when service rates are class independent)** Suppose $\mu_i = \mu$, $1 \leq i \leq K$, then
(i) the limiting frontier process \( \hat{H} \) satisfies the one-dimensional OU type SDE

\[
\eta(t)\hat{H}(t) = -\int_0^t \eta(u)\hat{H}(u)du + S(t) + G(t),
\]

(2.37)

where \( G(t) \equiv \sum_{i=1}^K G_i(t) \), for \( G_i(t) \) being the Brownian-driven terms given in Theorem 2.3.1, and

\[
S(t) \equiv \sum_{i=1}^K \eta_i(t)\kappa_i(t) + \int_0^t \eta_i(u)h_{F_i}(w_i)\kappa_i(u)du - c(t) - \mu \int_0^t c(u)du.
\]

(ii) The SDE (A.14) has a unique solution

\[
\hat{H}(t) = \frac{1}{R(t)} \left( \int_0^t e^{\int_u^t L(v)dv} \hat{f}(u)du + \int_0^t e^{\int_u^t L(v)dv} R(u)dK(u) \right)
+ \int_0^t e^{\int_u^t L(v)dv} K(u)dR(u),
\]

(2.38)

where \( \mathcal{W} \) is a standard Brownian motion,

\[
R(t) = e^{\mu t} \eta(t), \quad \tilde{L}(t) = e^{\mu t} \sum_{i=1}^K \eta_i(t) (\mu - h_{F_i}(w_i)),
\]

\[
\tilde{f}(t) = e^{\mu t} \sqrt{\sum_{i=1}^K (F_i^c(w_i)\lambda_i(t-w_i) + \mu m_i(t))}.
\]

(iii) The variance of \( \hat{H}(t) \) is

\[
\sigma_{\hat{H}}^2(t) \equiv \text{Var}(\hat{H}(t)) = \frac{1}{R^2(t)} \int_0^t e^{2\int_u^t L(v)dv} \tilde{f}^2(u)du.
\]

(2.39)

We next consider some special cases to obtain insights.

**Corollary 2.3.3 (Constant arrival rates)** When \( \lambda_i(t) = \lambda_i \), we have

\[
m_i(t) \sim m_i \equiv \frac{\lambda_i F_i^c(w_i)}{\mu}, \quad c(t) \sim c \equiv \sum_{i=1}^K \frac{w_i \lambda_i f_i(w_i)}{\mu} \kappa_i,
\]

(2.40)

\[
\kappa_i(t) \sim \kappa_i \equiv \sqrt{\sum_{j=1}^K \frac{\lambda_j F_j^c(w_j)}{\sum_{j=1}^K \lambda_j f_j(w_j) w_j}}.
\]

(2.41)

where we say \( f(t) \sim g(t) \) if \( f(t)/g(t) \to 1 \) as \( t \to \infty \).
Remark 2.3.6 (Average staffing and prioritization levels) The constants in (2.40) and (A.24) can be used to compute the required average number of servers and scheduling threshold. When \( K = 1 \), our staffing formula (2.40) degenerates to the ED+QED staffing formulas (30) and (31) in [73] which asymptotically controls the TPoD for the stationary \( M/M/n+G \) model.

In addition, these analytic formulas can provide an estimate of the marginal prices of staffing and scheduling (MPSS), that is, to improve the service to the next level (e.g., reducing \( w_i \) by \( \Delta w_i \), or reducing \( \alpha_i \) by \( \Delta \alpha_i \)), how many extra servers are need and how much should the scheduling threshold \( \kappa_i \) be adjusted?

If \( K = 1 \), then our multiclass \( V \) model degenerates to a single-class \( M_t/M/s_t+GI \) model.

Corollary 2.3.4 (The single-class case) When \( K = 1 \), the second-order staffing term \( c(t) \) simplifies to

\[
c(t) = z_1 e^{-\alpha t} \left( Z(t) - (\mu - h(w)) \int_0^t Z(s) ds \right),
\]

with

\[
Z(t) = e^{(\mu - h(w))t} \sqrt{\int_0^t e^{2h(w)} \left( F^c(w)\lambda(u-w) + \mu m(u) \right)}.
\]

It is easy to check that (2.42) and (2.43) coincide with the staffing formulas (7) and (8) in [62], except for a time shift by \( w \). This is due to the slightly different initial condition here.

2.4 Numerical Studies

In this section, we provide numerical examples and simulation comparisons to test the effectiveness of our TV-SRS and TV-DPS formulas. In §2.4.1 we first consider a base model having two customer classes and state-independent service rates. We next give additional simulation experiments in §2.4.2, including cases with smaller arrival rates and number of servers, higher quality of service, mixed scales of arrival rates, state-dependent service rates, and a five-class example.

2.4.1 A Two-Class Base Model

Because sinusoidal functions capture the periodic structure in realistic arrival patterns (see [20, 29, 66]), we consider sinusoidal arrival rates

\[
\lambda_i(t) = \bar{\lambda}_i \left( 1 + r_i \sin(\gamma_i t + \phi_i) \right), \quad 1 \leq i \leq K,
\]

with average rate \( \bar{\lambda}_i \), relative amplitude \( |r_i| < 1 \), frequency \( \gamma_i \), and phase \( \phi_i \). Corollary 2.4.1 provides a simplified first-order staffing term under this sinusoidal assumption.
Corollary 2.4.1 If arrival rates are sinusoidal as in (2.44), then

\[ m_i(t) \sim \frac{\bar{\lambda}_i F_c^i(w_i)}{\mu} + \frac{r_i \bar{\lambda}_i F_c^i(w_i)}{\sqrt{\mu^2 + \gamma_i^2}} \sin(\gamma_i t + \phi_i - \psi_i), \quad \psi_i \equiv \arctan\left(\frac{\gamma_i}{\mu}\right). \]  

We first consider a two-class V model, where Class 1 and Class 2 represent high and low priority customers respectively. We let \( \bar{\lambda}_1(t) = 1, \bar{\lambda}_2(t) = 1.5, r_1 = 0.2, r_2 = 0.3, \gamma_1 = \gamma_2 = 1, \phi_1 = 0, \phi_2 = -1 \). Abandonment times follow class-dependent exponential distributions with PDF \( f_i(x) = \theta_i e^{-\theta_i x} \).

We let \( \theta_1 = 0.6, \theta_2 = 0.3 \). Service rates are class-independent and standardized so that \( \mu_1 = \mu_2 = 1 \) (with mean service time \( 1/\mu_i = 1 \)). To prioritize Class 1, we set higher QoS levels (i.e., lower target wait time and tail probability of delay). We set our target model parameters as \( w_1 = 0.5, w_2 = 1, \alpha_1 = 0.2, \alpha_2 = 0.8 \).

Figure 2.3: Computed control functions for a two-class base case: \( m(t), c(t), \kappa_i(t) \) and \( \sigma(t), i = 1, 2 \).

In Figure 2.3, we calculate and plot the required control functions for TV-SRS and TV-DPS in a finite time interval \([0, T]\), with \( T = 24 \), including the offered-load function \( m(t) \) in (2.10), the second-order staffing term \( c(t) \) in (A.9), the second-order prioritization regulators (A.10), and the standard deviation process of \( \hat{H} \) in (A.15). Consistent with discussions in Remarks 2.2.2 and 2.3.4, we observe that \( \kappa_1(t) > 0 \) and \( \kappa_2(t) < 0 \) because \( \alpha_1 = 0.2 < 0.5 < 0.8 = \alpha_2 \). In addition, the second-order safety
staffing term, \( c(t) \), can be alternating between positive and negative. This is because the second order is a refinement term and indicates the service capacity that must be added or removed from the first order capacity in order to meet TPoD requirements.

![Figure 2.4: Simulation comparison for a two-class base case: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD \( P(V_i(t) > w_i) \) (middle panel); and (iii) time-varying staffing level (bottom panel), with \( w_1 = 0.5 \), \( w_2 = 1 \), \( \alpha_1 = 0.2 \), \( \alpha_2 = 0.8 \), and 5000 independent runs.](image)

Using these control functions in Figure 2.3, we conduct Monte-Carlo simulation experiments to test the effectiveness of TV-SRS and TV-DPS. For our base case, we let \( n = 50 \) and generate 5000 independent runs. Specifically, at each time \( 0 \leq t \leq T \) on an arbitrary run, we schedule the next customer into service according to our TV-DPS in (2.13) using the control function \( \kappa_i \) given in Figure 2.3. We plot (i) arrival rates, (ii) simulations of TPoD, and (iii) staffing functions, in Figure 2.4, using a sampling resolution (i.e., step size) \( \Delta t = 0.01 \). From a visual inspection of the middle panel of Figure 2.4, we see that our method effectively achieves stabilization of TPoD \( P(V_i(t) > w_i) \) for both classes at their (differentiated) targets (dashed lines). Implementation details of the simulations are discussed in §2.4.3.

**Staffing Discretization**

In practice (and in our simulation experiments), our TV-SRS formula needs to be discretized to integer values. Table 2.1 gives the time-averaged, maximum, and minimum simulation results for the two
TPoD’s and their relative differences from targets \(\mathbb{P}(V_i(t) > w_i) - \alpha_i)/\alpha_i\), using three staffing discretization methods (flooring, rounding, ceiling).

Table 2.1 exhibits the impact of adding and removing a server on the TPoD performance. As shown in the table, the discretization method seems to play a bigger role when the target QoS is high (\(\alpha\) is small), as in the case of Class 1. There is approximately a 20% difference in the time-average of TPoD for Class 1 between the flooring and ceiling methods. This highlights the impact a single server can have on TPoD control when target QoS is high, even when total service capacity is two magnitudes larger. In contrast, Class 2 with low QoS is relatively insensitive to the discretization method. All methods give average points falling within approximately \(\pm 5\%\) of the target. Interestingly, the rounding method does not yield the time-average closest to the target for Class 2, with the flooring method providing a closer in magnitude though greater than desired average. In general, to provide model performance stabilization closest to the target TPoD values, the rounding discretization method should be used. However, oftentimes QoS targets levels must be met or improved upon. In an effort to ensure time-average TPoD performance is less than or equal to the target QoS levels, we conduct the remainder of the simulations with the ceiling discretization method. As the scale \(n\) increases, the discretization becomes insignificant and all methods will provide nearly equivalent TPoD performance.

<table>
<thead>
<tr>
<th>Class</th>
<th>Avg.</th>
<th>Max</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Floor</td>
<td>Round</td>
<td>Ceiling</td>
</tr>
<tr>
<td>1</td>
<td>(\alpha_1)</td>
<td>0.2312</td>
<td>0.2091</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>(+15.59)</td>
<td>(+4.53)</td>
</tr>
<tr>
<td>2</td>
<td>(\alpha_2)</td>
<td>0.8085</td>
<td>0.7855</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>(+1.06)</td>
<td>(-1.81)</td>
</tr>
</tbody>
</table>

**Fixed staffing intervals**

In practice, system managers are often unable to add and remove servers in a nearly continuous manner; they must staff at a certain level for a fixed period of time (i.e. shifts in a hospital). We further expand upon the discretization of the continuous staffing function, by letting staffing decisions be limited to fixed intervals, in which the staffing levels must remain constants. We explore the impact of modifying our prescribed staffing formula to mimic this practical constraint. For a given staffing interval \(\Delta_s\) (e.g., 30 minutes) and a continuous TV-SRS formula \(s(t)\), we consider two \(\Delta_s\)-based discretization methods.
(i) **average staffing level (ASL)** and (ii) **maximum staffing level (MSL)**, which are given by

\[
\begin{align*}
    s_{\text{ASL}}(t) &\equiv \left\lceil \frac{T}{\Delta s} \right\rceil \sum_{i=1}^{\left\lceil T/\Delta s \right\rceil} \bar{s}_i \mathbb{1}_{\{t \in [(i-1)\Delta_s, i\Delta_s)\}}, \\
    s_{\text{MSL}}(t) &\equiv \left\lceil \frac{T}{\Delta s} \right\rceil \sum_{i=1}^{\left\lceil T/\Delta s \right\rceil} s^\uparrow_i \mathbb{1}_{\{t \in [(i-1)\Delta_s, i\Delta_s)\}},
\end{align*}
\]

where \( \bar{s}_i \equiv \frac{1}{\Delta s} \int_{(i-1)\Delta_s}^{i\Delta_s \wedge T} s(u) \mathrm{d}u \), \( s^\uparrow_i \equiv \sup_{(i-1)\Delta_s \leq u \leq i\Delta_s \wedge T} s(u) \), and \( x \wedge y \equiv \min(x, y) \). MSL sets the staffing level in each interval as the maximum of TV-SRS.

Figure 2.5: Simulation comparison for a two-class base case: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD \( \mathbb{P}(V_i(t) > w_i) \) (middle panel); and (iii) time-varying ASL and MSL staffing levels having \( \Delta_s = 0.5 \) (bottom panel), with \( w_1 = 0.5 \), \( w_2 = 1 \), \( \alpha_1 = 0.2 \), \( \alpha_2 = 0.8 \), and 5000 independent runs.

ensuring target QoS to be met as we will be slightly overstaffing the system; while ASL uses the average staffing level in each interval to ensure a smaller absolute deviation from the TPoD target. We again simulate our two-class base case model, but with staffing formulas calculated according to the ASL and MSL methods. We give our simulation results with \( \Delta_s = 0.5 \) (30 minutes) in Figure 2.5. We observe that both ASL and MSL achieve relative performance stabilization after an initial warm-up period (approximately the interval \([0, 4]\)). During the warm-up period, the rate of change in the required staffing is high and an inflexible staffing interval is not able to respond dynamically enough to meet demand. As a result, we have the greatest deviation from target TPoD levels occurring during this warm-up period. Indeed, ASL achieves better stabilization around the targets while MSL ensures meeting service levels at all times, leading to higher QoS than required.
2.4.2 Other Cases

To provide more engineering insights, we next test the robustness of TV-SRS and TV-DPS by considering variates of the base model, including (i) higher QoS targets (§2.4.2), (ii) smaller system scale (§2.4.2), (iii) mixed class scales (§2.4.2), (iv) class-dependent service rates (§2.4.2), and (v) a five-class example (§2.4.2).

Higher QoS targets

In our base model, we set $\alpha_1 = 0.2$ and $\alpha_2 = 0.8$ to test if TPoDs can be indeed significantly differentiated. We now validate the effectiveness of TV-SRS and TV-DPS when both classes have higher QoS targets.

![Figure 2.6: The two-class based model with high QoS targets $\alpha_1 = 0.05$, $\alpha_2 = 0.1$.](image)

Figure 2.6 gives the simulation results with smaller probability targets $\alpha_1 = 0.05$ and $\alpha_2 = 0.1$ ($w_1 = 0.5$, $w_2 = 1$) and shows that TPoD’s remain relatively stable. The bigger stochastic fluctuations are attributed to the much finer scale of performance.

Class 2 achieves stability at a value less than the target $\alpha_2 = 0.1$. This can be attributed to the ceiling discretization method. By ceiling the staffing function, we are providing an extra server at every time step with a fractional server. This creates a surplus of service capacity needed to achieve target stabilization, thus resulting in less abandonment and lower TPoD. If there was an imbalance in scheduling
policy but the correct staffing, we would see an increase in one and decrease in another.

**Smaller arrival rates**

Our method is based on asymptotic analysis of the V model when \( n \to \infty \), so it is evident that we should be able to achieve desired TPoD performance when \( n \) is relatively large. An important question is how effective TV-SRS and TV-DPS are for a small-scale system. We now consider the two-class base model having a smaller scale \( n = 5 \). Due to the increased stochastic variability in small-scale models, we increase our sample size to 20000 independent runs in our Monte-Carlo simulations. Figure 2.7 shows

![Arrival Rate](image1)

![Tail Probability of Delay](image2)

![Staffing Level](image3)

Figure 2.7: Simulation comparison for a small-scale two-class model: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD \( P(V_i(t) > w_i) \) (middle panel); and (iii) time-varying staffing level (bottom panel), with \( n = 5, w_1 = 0.5, w_2 = 1, \alpha_1 = 0.2, \alpha_2 = 0.8, \) and 20000 independent runs, under three staffing discretizations.

that: (i) Due to the small arrival rate, the required staffing level is small, so that addition and removal of a single server from time to time lead to bigger TPoD bumps; (ii) Different staffing discretization methods now play bigger roles, that is, adding a server to the staffing level at all times can cause a much larger performance change; and nevertheless, (iii) our TV-SRS and TV-DPS yield relatively stable TPoD performance. This example shows that results derived from the large-scale (many-server) limits may have strong practical relevance, even for small-scale systems.
Mixed arrival rates

We look at the case where arrival rates are of different orders of magnitude. This is relevant because, in practice, certain customer classes may have infrequent arrivals as compared to other classes, see [27]. We modify the arrival rates in our two-class base case so that $\tilde{\lambda}_1 = 0.1$, but set $n = 100$ so that the overall system size remains comparable to the base case. We see from Figure 2.8 that, even though the majority of arrivals to the system are from Class 1, we have effective TPoD stabilization for both classes.

![Figure 2.8: Simulation comparison for a two-class model with mixed arrival rates: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ (middle panel); and (iii) time-varying staffing level (bottom panel), with $\tilde{\lambda}_1 = 0.1, \tilde{\lambda}_2 = 1.5, n = 100, w_1 = 0.5, w_2 = 1, \alpha_1 = 0.2, \alpha_2 = 0.8$, and 5000 independent runs.]

Mixed target delay

Next, we modify the base case by setting Class 1 and Class 2 target delay to 0.5 and 2, respectively. This models a customer of high priority that should wait relatively little time in queue as compared to the other classes. Figure 2.9 shows our method’s ability to meet both differentiated TPoD and delay targets, simultaneously. We have slight under performance in Class 1 with low target delay. Note that as $w_1 \to 0$, our model must approach the QED regime with respect to Class 1, where a Class 1 customer should be served immediately upon arrival. However, in order to achieve positive target wait times in Class 2, our prescribed service capacity is such that our model operates in the ED regime. To maintain
a positive queue length for Class 2, there are no idle servers. But, because both classes are served from a shared service pool, this means there are no idle servers to immediately serve Class 1 customers upon arrival. Thus, Class 1 customers will always wait in queue longer than their target wait time and, as a result, have a TPoD of 1. This presents a boundary case in our proposed method, as we are unable to achieve target TPoD levels. Future work includes addressing the multiclass case where a high-priority class has target delay of zero.

Class-dependent service rates

Results in §2.3 enables us to treat the case of class-dependent service rates, which has strong practical relevance. Taking the CTAS example, a patient of higher acuity may require a much longer treatment, resulting in a smaller service rate. We now consider our two-class base model with \( \mu_1 = 0.5 \) and \( \mu_2 = 1 \) (so that a high priority class requires significantly more time in service). To obtain the control parameters, we calculated \( \text{Var}(\hat{H}(t)) \) according to the contraction-based algorithm given in Remark 2.5.2. Simulations show that our methods continue to achieve desired service-level differentiation and performance stabilization. The performance figure is similar to Figure 2.4 (see Figure 2.12). We discuss the implementation and convergence behavior of our algorithm for \( \text{Var}(\hat{H}(t)) \) in Remark 2.5.2.
A Five-Class Example

Finally, motivated by the CTAS example, we now consider a five-class V model, having class-dependent sinusoidal arrival rates as in (2.44), exponential abandonment and service times. All model input parameters and QoS parameters are given in Table 2.3. The control functions are given in the left-hand panel of Figure 2.10. In this example, we intentionally let the sinusoidal arrival rates have class-dependent periods, frequencies, and relative amplitudes (see right-hand panel of Figure 2.10). Nevertheless, our method continues to successfully achieve stable TPoD-based service levels across all 5 classes.

Table 2.2: Five Class Model: Class specific parameters and QoS target levels for the TPoD stabilization model

<table>
<thead>
<tr>
<th>Class</th>
<th>( \lambda )</th>
<th>( r )</th>
<th>( \gamma )</th>
<th>( \phi )</th>
<th>( \theta )</th>
<th>( \mu )</th>
<th>( w )</th>
<th>( \alpha )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>0.20</td>
<td>0.5</td>
<td>0</td>
<td>0.6</td>
<td>1</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>0.30</td>
<td>1.0</td>
<td>-1</td>
<td>0.3</td>
<td>1</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>1.2</td>
<td>0.05</td>
<td>1.3</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>1.1</td>
<td>0.15</td>
<td>1.6</td>
<td>-2</td>
<td>1.0</td>
<td>1</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>5</td>
<td>1.6</td>
<td>0.40</td>
<td>2.0</td>
<td>2</td>
<td>1.2</td>
<td>1</td>
<td>1.0</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Figure 2.10: A five-class based model: (i) Computed control functions \( m(t), c(t), \) and \( \kappa_i(t) \) for \( i = 1, \ldots, 5 \) (left), (ii) Simulation comparisons for TPoD \( P(V_i(t) > w_i) \), \( i = 1, \ldots, 5 \) (right), with \( n = 50 \), input and QoS parameters given in Table 2.3, and 5000 samples.
2.4.3 Implementation Details

All Monte Carlo simulations were conducted using MATLAB. We sample the values of the performance functions at fixed time points $\Delta T, 2\Delta T, \ldots, N\Delta T = T$ where $T = 24$ is the length of the time interval, the step size (sampling resolution) is $\Delta T = 0.01$, and $N = T/\Delta T = 2400$ is the total number of samples in $[0, T]$. To collect simulated data of PWT, on each simulation run, we create frequent virtual arrivals at all queues with interarrival time $\Delta T$. These virtual customers behave like real customers while in the queue and capture what the system experience would be like for a customer had they arrived at the given sampling time points. However, these virtual customers, when they are eventually moved to the head of the queue and assigned with a server, will not enter service; instead, they are removed immediately from the system after their elapsed waiting times have been recorded. For instance, the $j^{th}$ $(1 \leq j \leq N)$ class-$i$ virtual customer arrives at queue $i$ at time $j\Delta T$. If this customer is removed (from the head of the line) at time $t$, then the system collects a sample for the class-$i$ PWT at time $j\Delta T$ on the $l^{th}$ run: $V^l_i(j\Delta T) = t - j\Delta T$. The class-$i$ mean PWT and TPoD at time $t_j \equiv j\Delta T$ are estimated by averaging $m$ (e.g., $m = 5000$) independent copies of $V^l_i(j\Delta T)$ and indicators $1\{V^l_i(j\Delta T) > w_i\}$, namely, we use the unbiased Monte-Carlo estimators

$$E[V_i(t_j)] \equiv \frac{1}{m} \sum_{l=1}^{m} V^l_i(j\Delta T) \quad \text{and} \quad P(V_i(t_j) > w_i) \equiv \frac{1}{m} \sum_{l=1}^{m} 1\{V^l_i(j\Delta T) > w_i\}.$$ 

The numerical integrations (for the variance formulas and control functions) were done using the trapezoidal method in MATLAB.

2.4.4 Additional Numerical Examples

Class-dependent service rates

We extend the two-class base model in §2.4.1 to the case of class-dependent service rates, $\mu_1 = 0.5$ and $\mu_2 = 1$. In this case we numerically compute the variance process of $\tilde{H}(t)$ and required control functions using our contraction based algorithm given in Remark 2.5.2. We pick the error tolerance $\epsilon = 10^{-6}$. It took our contraction-based algorithm 42 iterations to converge. Similar to the case of class-independent service rates, TV-SRS and TV-DPS continue to achieve good TPoD performance. See Figure 2.12 for the simulation results.

Bigger staffing intervals

We extend our discussion in §2.4.1 to further study the impact of inflexible staffing functions. We do so by increasing $\Delta_s$. In Figure 2.11, we compare the TPoD performance for a system staffed according to the MSL method with $\Delta_s = 0.5$ and $\Delta_s = 2$. As expected, when increasing the interval length, the
staffing level becomes too rough to cope with the time-varying demand, thus unable to achieve desired performance stabilization.

Figure 2.11: Simulation comparison for a two-class base case: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD \( P(V_i(t) > w_i) \) (middle panel); and (iii) time-varying ASL and MSL staffing levels having \( \Delta_s = 0.5 \) (bottom panel), with \( w_1 = 0.5, w_2 = 1, \alpha_1 = 0.2, \alpha_2 = 0.8, \) and 5000 independent runs.

Relaxation of the assumption on class-dependent arrival times

Recall that at the beginning of §2.2.1, we assumed that each class-\( i \) arrival process \( A_i(t) \) begins at time \(-w_i < 0, \ 1 \leq i \leq K \). Such an assumption is imposed to facilitate the mathematical treatment. We now elaborate: Given class-dependent delay targets \( 0 < w_1 \leq w_2 \leq \cdots \leq w_K \) (without loss of generality we order them in increasing order), each class-\( i \) arrival process begins at a different (negative time) \(-w_i\), and class-\( i \) arrivals begin to occur earlier than class-\( j \) arrivals for \( i > j \). By time 0 at which we begin to serve all customers following TV-SRS and RV-DPS, we already have enough candidate customers, and more importantly, each class-\( i \) HOL customer is “old” enough (reaching the specific class-\( i \) delay target \( w_i\)). This provides a clean condition for us to implement TV-DPS. (Note that we do not serve any customer until time 0 because our TV-SRS \( s(t) = 0 \) for \( t < 0 \).) We now relax the different arrival time assumption by allowing customers of all classes to arrive at the same time \( t = 0 \). Without loss of generality, we can simply let all customer classes begin their arrival processes at time \(-w_K = \max_{1 \leq i \leq K} w_i < 0 \). (It suffices to later shift time by \( w_K \) units.)

Comparing the two assumptions, we now allow additional arrivals to occur before time 0 for classes 1, 2, \ldots, \( K - 1 \). In the \( n^{th} \) system, the new assumption adds approximately \( O(n(w_K - w_i)) \) initial customers to the \( i^{th} \) queue, \( 1 \leq i \leq K - 1 \).
Figure 2.12: Simulation comparison for a two-class model with class-dependent rates: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ (middle panel); and (iii) time-varying staffing level (bottom panel), with $\mu_1 = 0.5, \mu_2 = 1, n = 50, w_1 = 0.5, w_2 = 1, \alpha_1 = 0.2, \alpha_2 = 0.8$, and 5000 independent runs.

We first give a numerical example to illustrate the effect of adding additional customers at time 0 under the new assumption with the system operating under TV-SRS and TV-DPS based on the first assumption. In Figure 2.13, we reuse our two-class base example. The second plot shows that the additional initial customer can have significant impact on the TPoD performance; they cause an initial rampdown (understaffing), see the dotted-and-dashed lines. We believe this performance is still acceptable, because stable TPoD is achieved after some initial warm up time.

Nevertheless, we next discuss how to adjust our staffing and scheduling formulas to eliminate the initial TPoD bumps. To treat the new (more complex) assumption, we follow the discussion in Remark 2.2.3. We partition the negative interval $[-w_K, 0)$ into $K$ consecutive subintervals: $\mathcal{I}_1 \equiv [-w_K, -w_K + w_1), \mathcal{I}_2 \equiv [-w_K + w_2, -w_K + w_3), \ldots, \mathcal{I}_K \equiv [-w_K + w_{K-1}, 0)$. In the interval $\mathcal{I}_1$, we do not serve any customers. In $\mathcal{I}_2$, we act as if there is one customer class, namely, class 1 (we compute the TV-SRS formula using arrival rate $\lambda_1(t)1\{t \in \mathcal{I}_1\}$). In $\mathcal{I}_3$, we act as if there is two customer classes, namely class 1 and 2 (we compute the TV-SRS and TV-DPS using arrival rate $\lambda_1(t)1\{t \in \mathcal{I}_2\}$ and $\lambda_2(t)1\{t \in \mathcal{I}_1 \cup \mathcal{I}_2\}$). Using the same two-class example, we plot the simulated TPoD curves in Figure 2.13 under the refined policy; see the two solid lines in plot 2. This shows that our refinement on our “piecewise” TV-SRS and TV-DPS in consecutive intervals indeed achieves stable TPoD performance starting from the beginning (without needing a warmup period). See plot 3 of Figure 2.13 for the two
staffing functions.

Figure 2.13: Simulation comparison for a two-class model with both classes begin to arrive at the same time \(-w_2\); (i) arrival rates (top panel); (ii) simulated class-dependent TPoD \(P(V_i(t) > w_i)\) under both staffing functions (middle panel); and (iii) time-varying staffing level (bottom panel), with \(\mu_1 = 0.5, \mu_2 = 1, n = 50, w_1 = 0.5, w_2 = 1, \alpha_1 = 0.2, \alpha_2 = 0.8\), and 5000 independent runs.

2.4.5 Staffing and Scheduling to Differentiate the Mean PWT

We have shown in Theorem 2 that TV-SRS and TV-DPS achieves asymptotic stabilization for both mean delay \(\mathbb{E}[V_i(t)]\) and TPoD \(P(H_i(t) > w_i)\) at desired targets \(w_i\) and \(\alpha_i\).

We will next provide simulations to confirm the effectiveness of our methods using \(\mathbb{E}[V_i(t)]\) as the performance metric to stabilize. First, we note that this is a performance metric that depends only on \(w_i\), not \(\alpha_i\); and our second-order control functions (second-order safety staffing term \(c\) and second-order scheduling threshold \(\kappa_i\)) are developed to account for the probability target \(\alpha_i\). Indeed, it is clear from Figure 2.14 and 2.18 that these second-order terms will play negligible roles as the scale \(n\) increases, which is consistent with our asymptotic stability results in Theorem 2. For simplicity, in this section we will set the probability target \(\alpha_i = 0\) for all \(1 \leq i \leq K\), so that \(c(t) = \kappa_i(t) = 0\) and TV-DPS and
TV-SRS degenerates to the simpler *HWT-based dynamic prioritization* and *offered-load staffing*

\[
i^* \in \{H_i(t) \over w_i\} \quad \text{and} \quad s(t) = m(t).
\] (2.46)

This simplification is intuitive: The basic idea of using the second-order terms to set the tail probability \( \mathbb{P}(V^n_i(t) > w_i) \) to \( \alpha_i \) is to asymptotically adjust the mean of the nearly Gaussian distributed \( \hat{V}^n_i(t) \) so that the probability mass above 0 is equal to \( \alpha_i \). According to the symmetric structure of the limiting Gaussian distribution, we should set \( \alpha_i = 0 \) in order to have the mean of \( V^n_i(t) \) balanced at \( w_i \).

Paralleling the set of numerical experiments conducted in the previous section, we consider the following cases:

(i) **Base case.** The two-class model described above with sinusoidal arrival rates in (2.44), and scale \( n = 50 \). See Figure 2.14.

(ii) **High QoS.** The example in (i) with smaller delay targets \( w_1 = 0.1, w_2 = 0.2 \). See Figure 2.15.

(iii) **Smaller arrival rate.** The example in (i) having the sinusoidal arrival rates in (2.44) with \( n = 5 \). See Figure 2.16.

(iv) **Mixed arrivals.** The example in (i) with class-1 arrival rate reduced by an order of magnitude (i.e., \( \bar{\lambda}_1 = 0.1 \)). See Figure 2.17.

(v) **Five-class example.** The five-class example in §2.4, with \( \alpha_i = 0.5, 1 \leq i \leq 5 \). See Figure 2.18.

<table>
<thead>
<tr>
<th>Class</th>
<th>( \bar{\lambda} )</th>
<th>( r )</th>
<th>( \gamma )</th>
<th>( \phi )</th>
<th>( \theta )</th>
<th>( \mu )</th>
<th>( w )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>0.20</td>
<td>0.5</td>
<td>0.6</td>
<td>1</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>0.30</td>
<td>1.0</td>
<td>0.3</td>
<td>1</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.2</td>
<td>0.05</td>
<td>1.3</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>4</td>
<td>1.1</td>
<td>0.15</td>
<td>1.6</td>
<td>-2</td>
<td>1.0</td>
<td>1</td>
<td>0.8</td>
</tr>
<tr>
<td>5</td>
<td>1.6</td>
<td>0.40</td>
<td>2.0</td>
<td>1.2</td>
<td>1</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

The Monte-Carlo simulations are conducted by generating 5000 independent runs.

### 2.5 Proofs

Here we prove Theorems 2.3.1–2.3.2, Propositions 1–2, and Corollaries 2.3.2, 2.3.3, and 2.3.4.
Because we assume each baseline arrival-rate function $\lambda_i$ is bounded away from zero and infinity, we define

$$\lambda_i^\downarrow \equiv \inf_{0 \leq t \leq T} \lambda(t) > 0 \quad \text{and} \quad \lambda_i^\uparrow \equiv \sup_{0 \leq t \leq T} \lambda(t) < \infty.$$ 

### 2.5.1 Proof of Theorem 2.3.1.

The proof proceeds in four major steps, as indicated by the proof sketch.

**Step 1: SSC for the pre-limit HWT and PWT processes.** We start by observing the relation between $H^n_i(t)$ and $V^n_i(t)$

$$V^n_i(t - H^n_i(t)) = H^n_i(t) + O(1/n) \quad (2.47)$$

under the TV-DPS rule, where the error term $O(1/n)$ will follow if the number of customers (from other queues) who have a higher service priority over the HoL customer in the $i$th queue is of order $O(1)$; i.e., it only requires $O(1)$ number of service completions before the HoL customer of the $i$th queue enters service. To see that relation (2.47) is true, suppose customer $A$ enters service from the $i$th queue at time $t$ and customer $B$ becomes the new HoL customer in queue $i$. Then customer $B$ must have
arrived at the system at time $t - H^n_i(t)$ Further we use $a^n_i$ to denote the inter-arrival time between $A$ and $B$. It is immediate that customer $A$ arrived at the system at time $t - H^n_i(t) - a^n_i$. Suppose $\kappa_i \equiv 0$, $i \in I \equiv \{1, \ldots, K\}$ (the case where $\kappa_i$ are not zero functions can be analyzed in a similar fashion). Then under the TV-DPS policy, only those class-$j$ customers who arrived during the interval

$$
\left( t - \frac{w_j (H^n_i(t) - a^n_i)}{w_i}, t - \frac{w_j (H^n_i(t))}{w_i} \right)
$$

(2.48)
could enter service prior to the time at which customer $B$ enters service. To proceed, we make the following observation: If $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ are two independent Poisson processes with rate $\lambda^{(1)}$ and $\lambda^{(2)}$, respectively, then the number of arrivals from $\mathcal{P}^{(2)}$ between two successive arrivals of $\mathcal{P}^{(1)}$ follows a geometric distribution with parameter with parameter $\frac{\lambda^{(1)}}{\lambda^{(1)} + \lambda^{(2)}}$. Now because the interval (2.48) has a length of $(w_j a^n_i / w_i)$, the number class-$j$ customer who can enter service before $B$ is stochastically bounded by a geometric distributed random variable with mean $\frac{w_j \lambda_j}{w_i \lambda_i}$ and variance $\left( \frac{w_j \lambda_j}{w_i \lambda_i} \right)^2 + \frac{w_j \lambda_j}{w_i \lambda_i}$. By the same token, we can argue that the total number of customers who will enter service before $B$ enters service is bounded by the sum of $K - 1$ geometric random variables with mean $\ell^{(1)}_i = \sum_{j \neq i} w_j \lambda_j / w_i \lambda_i$ and variance $\ell^{(2)}_i = \sum_{j \neq i} \left( \frac{w_j \lambda_j}{w_i \lambda_i} \right)^2 + \ell^{(1)}_i$.
On the other hand, the inter-arrival times of class $i$ live on the order of $O(1/n)$. Using the same reasoning for (2.47), we have

$$H^n_i(t)/w_i + n^{-1/2} \kappa_i(t) = H^n(t) - O(1/n),$$

or equivalently,

$$\hat{H}^n_i(t) = w_i(\hat{H}^n(t) - \kappa_i(t)) - O(1/\sqrt{n}),$$

(2.49)

where we recall that $\hat{H}^n$ is the CLT-scaled frontier process, namely, $\hat{H}^n(t) \equiv n^{1/2} (H^n(t) - 1)$.

**Step 2: The FWLLN.** Here we prove the desired FWLLN results by showing the stochastic boundedness of the corresponding CLT-scaled processes; see §5.2 of [77] for a precise definition of stochastic boundedness. In what follows, we will first prove that the sequence $\{(\hat{B}_1^n, \ldots, \hat{B}_K^n, \hat{H}_i^n); n \in \mathbb{N}\}$ is
Figure 2.17: Simulation comparison for a two-class model with mixed magnitudes of arrivals: (i) arrival rates (top panel); (ii) simulated class-dependent mean PWT $E[V_i(t)]$ (middle panel); and (iii) time-varying MSL staffing (bottom panel), with $\mu_1 = 0.5$, $\mu_2 = 1$, $n = 100$, $\bar{\lambda}_1 = 0.1$, $w_1 = 0.1$, $w_2 = 0.2$, and 5000 independent runs.

stochastically bounded. To that end, introduce the LLN- and CLT-scaled empirical process

$$
\hat{U}^n(t, x) = \frac{1}{n} \sum_{k=1}^{[nt]} 1\{X_i \leq x\} \quad \text{for} \quad t \geq 0, \quad 0 \leq x \leq 1, \quad \text{and}
$$

$$
\widehat{U}^n(t, x) = \sqrt{n} \left( \hat{U}^n(t, x) - E[\hat{U}^n(t, x)] \right) = \frac{1}{\sqrt{n}} \left( \sum_{k=1}^{[nt]} 1\{X_i \leq x\} - x \right),
$$

(2.50)

where $X_1, X_2, \ldots$ are i.i.d. random variables uniformly distributed on $[0, 1]$. [59] have shown that $\hat{U}^n \Rightarrow \widehat{U}$ in $D_D$ as $n \to \infty$, where $\widehat{U}$ is the standard Kiefer process. Paralleling (3.3) - (3.6) in [5], we break the enter-service process $E^n_i(t)$ in (2.6) into three pieces, namely,

$$
E^n_i(t) = E^n_{i,1}(t) + E^n_{i,2}(t) + E^n_{i,3}(t),
$$

(2.51)
Figure 2.18: Simulation comparison for a five-class model: (i) arrival rates (top panel); (ii) simulated class-dependent mean PWT $E[V_i(t)]$ (middle panel); and (iii) time-varying MSL staffing (bottom panel), with 5000 independent runs, and model parameters given in Table 2.3

where

$$E_{i,1}^n(t) \equiv \sqrt{n} \int_{-H_i^n(0)}^{t-H_i^n(t)} F_i^c(V_i^n(u))d\tilde{A}_i^n(u), \quad t \geq 0,$$

$$E_{i,2}^n(t) \equiv \sqrt{n} \int_{-H_i^n(0)}^{t-H_i^n(t)} \int_0^1 1_{\{y>F_i^c(V_i^n(u))\}}d\tilde{U}_i^n(\tilde{A}_i^n(u), y), \quad t \geq 0,$$

$$E_{i,3}^n(t) \equiv n \int_{-H_i^n(0)}^{t-H_i^n(t)} F_i^c(V_i^n(u))\lambda_i(u)du, \quad t \geq 0,$$

for $\tilde{A}_i^n$, $\tilde{A}_i^n$ given by (2.16) and $\tilde{U}_i^n$ is a CLT-scaled empirical process specified by (2.50).

Define the fluid version and CLT-scaled version of the enter-service process as

$$\varepsilon_i(t) \equiv \int_{-w_i}^{t-w_i} F_i^c(w_i)\lambda_i(u)du,$$

$$\tilde{E}_i^n(t) \equiv n^{-1/2}(E_i^n(t) - n\varepsilon_i(t)) = n^{-1/2}\left(E_i^n(t) - n \int_{-w_i}^{t-w_i} F_i^c(w_i)\lambda_i(u)du\right).$$
Following the decomposition given in (2.51) - (2.54), we can write
\[ \hat{E}_i^n(t) = \hat{E}_i^{n,1}(t) + \hat{E}_i^{n,2}(t) + \hat{E}_i^{n,3}(t), \] (2.57)
where
\[ \hat{E}_i^{n,1}(t) \equiv n^{-1/2} E_i^{n,1}(t) = \int_{-H_i^n(0)}^{t-H_i^n(t)} F_i^c(V_i^n(u))d\hat{A}_i^n(u) \quad t \geq 0 \] (2.58)
\[ \hat{E}_i^{n,2}(t) \equiv n^{-1/2} E_i^{n,2}(t) = \int_{-H_i^n(0)}^{t-H_i^n(t)} \int_0^1 1_{\{y>F_i^c(V_i^n(u))\}}d\hat{U}_i^n(A_i^n(u), y) \quad t \geq 0, \] (2.59)
\[ \hat{E}_i^{n,3}(t) \equiv n^{-1/2} \left( E_i^{n,3}(t) - n \int_{-w_i}^{t-w_i} F_i^c(w_i)\lambda_i(u)du \right) \quad t \geq 0. \] (2.60)

For the term \( \hat{E}_i^{n,3} \), we further deduce
\[
\hat{E}_i^{n,3}(t) = \sqrt{n} \left[ \int_{-H_i^n(0)}^{t-H_i^n(t)} F_i^c(V_i^n(u))\lambda_i(u)du - \int_{-w_i}^{t-w_i} F_i^c(w_i)\lambda_i(u)du \right]
\]
\[ = \sqrt{n} \int_0^t F_i^c(H_i^n(u))\lambda_i(u-H_i^n(u))du - \sqrt{n} \int_0^t F_i^c(w_i)\lambda_i(u-w_i)du
\]
\[ - \int_0^t F_i^c(H_i^n(u))\lambda_i(u-H_i^n(u))d\hat{H}_i^n(u) + O(n^{-1/2})
\]
\[ = - \int_0^t \left\{ f_i(\zeta_i^n(u))\lambda_i(u-\zeta_i^n(u)) + F_i^c(\zeta_i^n(u))\lambda_i'(u-\zeta_i^n(u)) \right\} w_i(\hat{H}_i^n(u) - \kappa_i(u))du
\]
\[ - \int_0^t w_i F_i^c(H_i^n(u))\lambda_i(u-H_i^n(u))d(\hat{H}_i^n(u) - \kappa_i(u)) + O(n^{-1/2}), \]
(2.61)
where the second equality follows by a change of variables, namely \( t \rightarrow t - H_i^n(t) \), plus the relation (2.47), while the third equality follows from (2.49) and applying the mean-value theorem with \( \zeta_i^n(t) \) satisfying
\[ H_i^n(t) \land w_i \leq \zeta_i^n(t) \leq H_i^n(t) \lor w_i. \] (2.62)

On the other hand, the conservation of flow implies
\[ E_i^n(t) = B_i^n(t) + D_i^n(t), \] (2.63)
where we have used \( D_i^n(t) \) to denote the number of class-\( i \) customers that have completed service by time \( t \). From (2.55) it follows
\[ \varepsilon_i(t) = \int_{-w_i}^{t-w_i} F_i^c(w_i)\lambda_i(u)du = \int_0^t F_i^c(w_i)\lambda_i(u-w_i)du = m_i(t) + \int_0^t \mu_i m_i(u)du, \] (2.64)
where the last equality follows from (2.10). Multiplying both sides of (2.64) by \( n \) and subtracting it from (2.63) yields
\[
E^n_i(t) - n\varepsilon_i(t) = B^n_i(t) - nm_i(t) + D^n_i(t) - n\int_0^t \mu_i m_i(u)du
\]
\[
= B^n_i(t) - nm_i(t) + \left(D^n_i(t) - \mu_i \int_0^t B^n_i(u)du\right)
+ \mu_i \left(\int_0^t B^n_i(u)du - \int_0^t nm_i(u)du\right).
\]

Next divide both sides by \( n^{1/2} \) to get
\[
\hat{E}^n_i(t) = \hat{B}^n_i(t) + \mu_i \int_0^t \hat{B}^n_i(u)du + \hat{D}^n_i(t) \quad \text{or} \quad d\hat{B}^n_i(t) + \mu_i \hat{B}^n_i(t)dt = d\hat{E}^n_i(t) - d\hat{D}^n_i(t),
\]
where we have defined
\[
\hat{D}^n_i(t) \equiv n^{-1/2} \left(D^n_i(t) - \mu_i \int_0^t B^n_i(u)du\right).
\]
Because the baseline input exceeds the output at all times, it holds that
\[
B^n_i(t) = s^n(t)
\]
with arbitrarily high probability for \( n \) sufficiently large. Hence, it suffices to focus on the sample paths for which relation (2.66) holds. In this case we can easily deduce
\[
\sum_{i=1}^K \hat{B}^n_i(t) = n^{-1/2} (B^n(t) - nm(t)) = n^{-1/2} (s^n(t) - nm(t)) = c(t).
\]
Upon substituting (2.57) - (2.59) and (2.61) into (2.65), we obtain, for \( i = 1, \ldots, K \),
\[
d\tilde{B}^n_i(t) + w_i F_i^c(H^n_i(t))\lambda_i(t - H^n_i(t))d\tilde{H}^n_i(t)
= -\mu_i \tilde{B}^n_i(t)dt - \left[f_i(\zeta^n_i(t))\lambda_i(t - \zeta^n_i(t)) + F_i^c(\zeta^n_i(t))\lambda'_i(t - \zeta^n_i(t))\right] w_i \tilde{H}^n_i(t)dt
+ \left[f_i(\zeta^n_i(t))\lambda_i(t - \zeta^n_i(t)) + F_i^c(\zeta^n_i(t))\lambda'_i(t - \zeta^n_i(t))\right] w_i \kappa_i(t)dt
+ w_i F_i^c(H^n_i(t))\lambda_i(t - H^n_i(t))d\kappa_i(t) + d\hat{E}^n_{i,1}(u) + d\hat{E}^n_{i,2}(u) - d\hat{D}^n_i(u) + O(n^{-1/2}).
\]
We can then use (2.67) to write \( \tilde{B}^n_K = c(t) - \sum_{i=1}^{K-1} \hat{B}^n_i \). Plugging it into (2.68) for \( i = K \), we obtain a set of \( K \) linear differential equations with respect to the \( K \)-dimensional process \((\tilde{B}^n_1, \ldots, \tilde{B}^n_K, \tilde{H}^n)\). Similar to what was done to (5.14) in [5], we apply the Gronwall’s inequality together with the stochastic boundedness of \( \tilde{E}^n_{i,1}, \tilde{E}^n_{i,2}, \text{and} \hat{D}^n_i \) plus the assumed properties of \( \lambda_i, f_i, \text{and} F_i^c \) to conclude the stochastic boundedness of the sequence \( \{B^n_1, \ldots, \hat{B}^n_{K-1}, \hat{H}^n; n \in \mathbb{N}\} \). In particular, the sequence
\(\{\hat{H}^n; n \in \mathbb{N}\}\) is stochastically bounded. In view of (2.49) and (2.47), we have that \(\{\hat{H}_n; n \in \mathbb{N}\}\) and \(\{\hat{V}_n; n \in \mathbb{N}\}\) are stochastically bounded, for \(i = 1, \ldots, K\). This implies the FWLLN for the HWT and PWT processes, that is, as \(n \to \infty\),

\[
(H^n, H^n_1, \ldots, H^n_K, V^n_1, \ldots, V^n_K) \Rightarrow (\varepsilon, w_1, \ldots, w_K, \varepsilon, w_1, \ldots, w_K) \quad \text{in} \quad \mathcal{D}^{2K+1},
\]

(2.69)

where the joint convergence holds due to converging-together lemma (Theorem 11.4.5. in [102]).

**Step 3: The FCLT for the waiting time processes.** Similar to the proof of Lemma 5.1 in [5], we invoke the continuous mapping theorem with (2.58) and (2.69) to get

\[
\hat{E}^n_{i,1}(t) \Rightarrow \hat{E}_{i,1}(t) \equiv F^c_i(w_i) \int_{-w_i}^{t-w_i} \lambda_i(u) dW_{\lambda_i}(u),
\]

(2.70)

where \(W_{\lambda_i}\) is a standard Brownian motion.

To proceed, we argue that, as \(n \to \infty\),

\[
\hat{E}^n_{i,2}(t) \Rightarrow \hat{E}_{i,2}(t) \equiv \sqrt{F^c_i(w_i)} F_i(w_i) \int_{-w_i}^{t-w_i} \lambda_i(u) dW_{\theta_i}(u),
\]

(2.71)

for \(W_{\theta_i}\) being a standard Brownian independent of \(W_{\lambda_i}\). The essential structure of the proof for (2.71) is exactly the same as that of A.7.2 in [5], which in turn draws on Theorem 7.1.4 in [28]. Because the proof can be fully adapted from theirs, we omit the details.

Moreover, as a direct consequence of the established stochastic boundedness of \(\{\hat{B}^n_1, \ldots, \hat{B}^n_K; n \in \mathbb{N}\}\), we have the FWLLN for the busy-server processes

\[
(\hat{B}^n_1, \ldots, \hat{B}^n_K) \Rightarrow (m_1, \ldots, m_K) \quad \text{in} \quad \mathcal{D}^K \quad \text{as} \quad n \to \infty.
\]

Next a standard random-time-change argument allows us to derive

\[
\hat{D}^n_i(\cdot) = n^{-1/2} \left[ \Pi^d_i \left( n\mu_i \int_0^{\hat{B}^n_i(u)} du \right) - n\mu_i \int_0^{\hat{B}^n_i(u)} du \right] \Rightarrow W_{\mu_i} \left( \mu_i \int_0^{m_i(u)} du \right), \quad n \to \infty,
\]

(2.72)

where we have defined \(\Pi^d_i\) to be a unit-rate Poisson process and \(W_{\mu_i}\) to be a standard Brownian motion independent of \(W_{\lambda_i}\) and \(W_{\theta_i}\). To establish the convergence of (2.19), we will need to strengthen (2.70), (2.71) and (2.72) to joint convergence. The joint convergence of multiple random elements is equivalent to individual convergence if they are independent. Here \(\hat{E}^n_{i,1}, \hat{E}^n_{i,2}\) and \(\hat{D}^n_i\) are not independent because both \(\hat{E}^n_{i,1}\) and \(\hat{E}^n_{i,2}\) involve the arrival-time sequence, and \(\hat{D}^n_i\) depends on \(B^n_i\) which in turn correlates with \(E^n_i\) through (2.63). But they are conditionally independent given \(A^n_i, H^n_i, V^n_i\) and \(B^n_i\). Hence, we can establish the joint convergence by first conditioning and then unconditioning. See Lemma 4.1 of [6]
for a reference, which is a variant of Theorem 7.6 of [77].

To derive a set of SDEs satisfied by the CLT-scaled processes \( (\hat{H}^n, \hat{B}^n_1, \ldots, \hat{B}^n_K) \), we seek to simplify the right-hand side of (2.61). First we note that the inequality (2.62) and the convergence in (2.71) imply

\[
\zeta^i(t) = w_i + O(n^{-1/2}) = H^i(t) + O(n^{-1/2}). \tag{2.73}
\]

We then use integration by parts to deduce

\[
- \int_0^t \left[ \sum_{i=1}^K w_i f^c_i(w_i) \lambda_i(t - \zeta^i(t)) (\hat{H}^n(t) - \kappa_i(t)) \right] dt
\]

\[
- \int_0^t \left[ \sum_{i=1}^K w_i F^c_i(\zeta^i(t)) \lambda_i(t - \zeta^i(t)) (\hat{H}^n(t) - \kappa_i(t)) \right] dt
\]

\[
- \int_0^t \left[ \sum_{i=1}^K \frac{F^c_i(\zeta^i(t))}{\zeta^i(t)} (\hat{H}^n(t) - \kappa_i(t)) \right] dt
\]

\[
= - w_i \{ F^c_i(\zeta^i(t)) \lambda_i(t - \zeta^i(t)) \} \frac{F^c_i(\zeta^i(t))}{\zeta^i(t)} \lambda_i(t - \zeta^i(t)) dt
\]

\[
+ O(n^{-1/2}), \tag{2.74}
\]

where the last equality holds due to (2.73). Upon plugging (2.74) into (2.61), we obtain

\[
\tilde{E}_{i,3}^n(t) = - \int_0^t \left[ \sum_{i=1}^K w_i f_i(w_i) \lambda_i(t - \zeta^i(t)) (\hat{H}^n(t) - \kappa_i(t)) \right] dt
\]

\[
- w_i \{ F^c_i(\zeta^i(t)) \lambda_i(t - \zeta^i(t)) \} \frac{F^c_i(\zeta^i(t))}{\zeta^i(t)} \lambda_i(t - \zeta^i(t)) dt + O(n^{-1/2}).
\]

Now plugging (2.57) and the equation above into (2.65), we get

\[
\tilde{B}_i^n(t) + w_i F^c_i(w_i) \lambda_i(t - w_i) \hat{H}_i^n(t) = - \mu_i \int_0^t \tilde{B}_i^n(u) du - \int_0^t \left[ \sum_{i=1}^K \frac{f_i(w_i)}{w_i} \lambda_i(t - w_i) \hat{H}_i^n(u) du \right]
\]

\[
+ \int_0^t \left[ \sum_{i=1}^K w_i f_i(w_i) \lambda_i(u - w_i) \kappa_i(u) du + w_i F^c_i(w_i) \lambda_i(t - w_i) \kappa_i(t) \right] dt
\]

\[
+ \tilde{E}_{i,1}^n(t) + \tilde{E}_{i,2}^n(t) - \tilde{D}_i^n(t) + O(n^{-1/2}) \quad \text{for } i = 1, \ldots, K. \tag{2.75}
\]

The joint convergence \( (\hat{H}^n, \hat{B}^n_1, \ldots, \hat{B}^n_K) \) \( \Rightarrow \) \( (\hat{H}, \hat{B}_1, \ldots, \hat{B}_K) \) then follows by applying the continuous mapping theorem (see Theorem 4.1 of [77]) to (2.66) and (2.75), with the joint convergence of \( \tilde{E}_{i,1}^n, \tilde{E}_{i,2}^n \) and \( \tilde{D}_i^n \), as specified by (2.70), (2.71) and (2.72), respectively. Alternatively, one can subtract (2.75) by (2.19) and invoke the Gronwall’s inequality to show that the difference between the pre-limit and the limit is bounded by a negligible term as \( n \to \infty \), as was done in the proof of (4.7) in [5]. The convergence of \( \{\hat{H}^n; n \in \mathbb{N}\} \) and \( \{\hat{V}^n; n \in \mathbb{N}\} \) follow easily from and (2.49) and (2.47), respectively.
Step 4: The FCLT for the queue-length processes. To show that \( \{ \hat{Q}_i^n; n \in \mathbb{N} \} \) converges to the corresponding limit, we decompose the right-hand side of (2.7) into three terms, namely,

\[
Q_i^n(t) = Q_{i,1}^n(t) + Q_{i,2}^n(t) + Q_{i,3}^n(t),
\]

where

\[
Q_{i,1}^n(t) \equiv \sqrt{n} \int_{t-H_i^n(t)}^t F_i^c(t-u) d\hat{A}_i^n(u), \quad t \geq 0, \tag{2.77}
\]

\[
Q_{i,2}^n(t) \equiv \sqrt{n} \int_{t-H_i^n(t)}^t \int_0^1 1_{\{x>F_i^c(t-u)\}} d\hat{U}_i^n(\hat{A}_i^n(u), x) \quad t \geq 0, \tag{2.78}
\]

\[
Q_{i,3}^n(t) \equiv n \int_{t-H_i^n(t)}^t F_i^c(t-u) \lambda_i(u) du \quad t \geq 0, \tag{2.79}
\]

Accordingly, the centered and normalized queue-length process can be decomposed into three terms

\[
\hat{Q}_i^n(t) \equiv n^{-1/2} (Q_i^n(t) - nq_i(t)) = \hat{Q}_{i,1}^n(t) + \hat{Q}_{i,2}^n(t) + \hat{Q}_{i,3}^n(t),
\]

where

\[
\hat{Q}_{i,1}^n(t) \equiv \int_{t-H_i^n(t)}^t F_i^c(t-u) d\hat{A}_i^n(u) \Rightarrow \int_{t-w_i}^t F_i^c(t-u) d\hat{A}_i(u), \tag{2.80}
\]

\[
\hat{Q}_{i,2}^n(t) \equiv \int_{t-H_i^n(t)}^t \int_0^1 1_{\{x>F_i^c(t-u)\}} d\hat{U}_i^n(\hat{A}_i^n(u), x) \Rightarrow \int_{t-w_i}^t \sqrt{F_i^c(t-u) F_i(t-u) \lambda_i(u) dW_i(u)}, \tag{2.81}
\]

\[
\hat{Q}_{i,3}^n(t) \equiv n \int_{t-H_i^n(t)}^{t-w_i} F_i^c(t-u) \lambda_i(u) du \Rightarrow F_i^c(w_i) \lambda_i(t-w_i) \hat{H}_i(t). \tag{2.82}
\]

Here the proof for (2.80) and (2.81) is very similar to that or (2.70) and (2.71), and the proof for (2.82) is also straightforward. □

2.5.2 Proof of Proposition 1.

The multi-dimensional SDE (2.19) is equivalent to

\[
\frac{d}{dt} \left( e^{\mu t} \hat{B}_i(t) \right) = e^{\mu t} \left( -w_i F_i^c(w_i) \lambda_i(t-w_i) \hat{H}(t) - \int_0^t w_i f_i(w_i) \lambda_i(u-w_i) \hat{H}(u) du + y_i(t) + G_i(t) \right), \tag{2.83}
\]
where

\[ \tilde{B}_i(t) \equiv \int_0^t \tilde{B}_i(u) du \quad \text{and} \quad y_i(t) \equiv w_i F_i^c \lambda_i (t - w_i) \kappa_i (t) + \int_0^t w_i f_i (w_i) \lambda_i (u - w_i) \kappa_i (u) du. \]

Integrating (A.7) from 0 to \( W \)

\[ \tilde{B}_i(t) = e^{-\mu_i t} \int_0^t e^{\mu_i s} \left( -w_i F_i^c \lambda_i(s - w_i) H(s) - \int_0^s w_i f_i (w_i) \lambda_i(u - w_i) H(u) du \right. \]

\[ + y_i(s) + G_i(s) \right) ds \]

\[ = e^{-\mu_i t} \left( -\int_0^t e^{\mu_i s} w_i F_i^c \lambda_i(s - w_i) H(s) ds - \int_0^t w_i f_i (w_i) \lambda_i(u - w_i) H(u) \int_u^t e^{\mu_i s} ds du \right. \]

\[ + \int_0^t e^{\mu_i s} y_i(s) ds + \int_0^t e^{\mu_i s} G_i(s) ds \right) \]

\[ = \int_0^t w_i \lambda_i(s - w_i) \left( -F_i^c(w_i) e^{\mu_i(s-t)} - f_i(w_i) \frac{1 - e^{\mu_i(s-t)}}{\mu_i} \right) H(s) ds \]

\[ + \int_0^t e^{\mu_i(s-t)} y_i(s) ds + \int_0^t e^{\mu_i(s-t)} G_i(s) ds. \]

Summing up over \( i \) from 1 to \( K \), we have

\[ \int_0^t c(s) ds = \sum_{i=1}^K \tilde{B}_i(t) = \int_0^t \sum_{i=1}^K w_i \lambda_i(s - w_i) \left( -F_i^c(w_i) e^{\mu_i(s-t)} - f_i(w_i) \frac{1 - e^{\mu_i(s-t)}}{\mu_i} \right) H(s) ds \]

\[ + \sum_{i=1}^K \int_0^t e^{\mu_i(s-t)} \left( w_i F_i^c(w_i) \lambda_i(s - w_i) \kappa_i(s) + \int_0^s w_i f_i (w_i) \lambda_i(u - w_i) \kappa_i(u) du \right) ds \]

\[ + \sum_{i=1}^K \int_0^t e^{\mu_i(s-t)} \left( \sqrt{F_i^c(w_i) \lambda_i(u - w_i)} + \mu_i m_i(u) dW_i(u) \right) ds \]

\[ = \sum_{i=1}^K \int_0^t w_i \lambda_i(s - w_i) \left( -F_i^c(w_i) e^{\mu_i(s-t)} - f_i(w_i) \frac{1 - e^{\mu_i(s-t)}}{\mu_i} \right) H(s) ds \]

\[ + \sum_{i=1}^K \int_0^t w_i \lambda_i(s - w_i) \kappa_i(u) \left( F_i^c(w_i) e^{\mu_i(s-t)} + f_i(w_i) \frac{1 - e^{\mu_i(s-t)}}{\mu_i} \right) du \]

\[ + \sum_{i=1}^K \int_0^t \frac{1 - e^{\mu_i(u-t)}}{\mu_i} \sqrt{F_i^c(w_i) \lambda_i(u - w_i)} + \mu_i m_i(u) dW_i(u), \]

where the second equality holds by aggregating three independent Brownian motions \( W_{\mu_i}, W_{\mu_i} \) and \( W_{\lambda_i} \) in (2.21) into one independent standard Brownian motion \( W_i \) for each \( 1 \leq i \leq K \). Differentiating
(2.84) yields
\[ c(t) = -\sum_{i=1}^{K} w_i \lambda_i (t - w_i) F_i^c(w_i) \tilde{H}(t) + \int_0^t \sum_{i=1}^{K} w_i \lambda_i (s - w_i) e^{\mu_i (s-t)} (\mu_i F_i^c(w_i) - f_i(w_i)) \tilde{H}(s) ds \]
\[ + \sum_{i=1}^{K} w_i \lambda_i (t - w_i) F_i^c(w_i) \kappa_i(t) + \int_0^t \sum_{i=1}^{K} w_i \lambda_i (s - w_i) e^{\mu_i (s-t)} (-\mu_i F_i^c(w_i) + f_i(w_i)) \kappa_i(s) ds \]
\[ + \sum_{i=1}^{K} \int_0^t e^{\mu_i (u-t)} \sqrt{F_i^c(w_i) \lambda_i (u - w_i) + \mu_i m_i(u)} dW_i(u). \]

And further aggregating the independent Brownian motions $\mathcal{W}_1, \ldots, \mathcal{W}_K$ into $\mathcal{W}$ yields the SVE in (2.26).

**Uniqueness and existence of solution to the SVE (2.26).** Consider two functions $x, y \in \mathbb{C}$ (space of continuous functions) satisfying an equation
\[ x(t) = \int_0^t L(t, s) x(s) ds + y(t). \tag{2.85} \]
we show that (2.85) specifies a well-defined function $\phi : \mathbb{C} \to \mathbb{C}$ such that $x = \psi(y)$. To do so, for a given $y$, we define the operator
\[ \psi(x)(t) \equiv \int_0^t L(t, s) x(s) ds + y(t). \tag{2.86} \]
Therefore, $x$ solves the fixed-point equation (FPE)
\[ x = \psi(x). \tag{2.87} \]

We first prove that $\psi$ is a contraction over a finite interval $[0, T]$. Specifically, let $x_1, x_2 \in \mathbb{C}$, and use the uniform norm $\|x\|_T = \sup_{0 \leq t \leq T} |x(t)|$. We have
\[ |\psi(x_1)(t) - \psi(x_2)(t)| \leq \int_0^t |L(t, s)| ds \cdot \|x_1 - x_2\|_T \]
\[ \leq \|x_1 - x_2\|_T \left( \sum_{i=1}^{K} w_i \lambda_i^\top (\mu_i F_i^c(w_i) + f_i(w_i)) \right) t. \tag{2.88} \]

Hence, we have $\|\psi(x_1) - \psi(x_2)\|_T \leq L^\top T \|x_1 - x_2\|_T$, where the constant
\[ L^\top = \frac{\sum_{i=1}^{K} w_i \lambda_i^\top (\mu_i F_i^c(w_i) + f_i(w_i))}{\sum_{i=1}^{K} w_i \lambda_i^\top F_i^c(w_i)} < \infty, \tag{2.89} \]
which is guaranteed by the strict positivity assumptions on \( w_i, \lambda_i \) and \( F_i^c \) for all \( 1 \leq i \leq K \). In case \( L^\uparrow T > 1 \), we can partition the interval \([0, T]\) to successive smaller intervals with length \( \Delta T \) satisfying \( \Delta T < 1/L^\uparrow \). This will recursively guarantee the contraction property over all smaller intervals. Hence, the Banach fixed point theorem implies that the FPE (2.87) has a unique solution over the entire interval \([0, T]\).

Consequently, the function \( \phi \) specified by (2.85) is well-defined because \( \phi(y) \) has one and only one image for any \( y \). So we conclude that (2.26) has a unique solution \( \tilde{H} \). If fact, we can write (2.26) as

\[
\tilde{H} = \phi \left( \int_0^t J(\cdot, s) dW(s) + K(\cdot) \right).
\]

**Remark 2.5.1** *The strict positivity assumptions on \( \lambda_i \) and \( F_i^c \) for all classes \( 1 \leq i \leq K \) can be relaxed. Note that the contraction property (2.88) continues to hold as long as there exists some (not all) class \( i \) such that \( \lambda_i^\uparrow \) and \( F_i^c(w_i) \) are both positive.*

To show that \( \tilde{H} \) is Gaussian, we again use the contraction \( \psi \) defined in (2.86). We follow the steps that establish strong solutions in [50]. Define a sequence of processes \( \{\tilde{H}^{(k)}, k = 0, 1, 2, \ldots\} \) such that \( \tilde{H}^{(0)}(t) = 0 \), and \( \tilde{H}^{(k+1)} = \psi(\tilde{H}^{(k)}) \) with \( y(t) = \int_0^t J(t, s) dW(s, \omega) \) for \( k \geq 0 \). (For each Brownian path and associated Brownian integral, we recursively define the sequence.) We can show that \( \tilde{H}^{(k)} \) is Gaussian using an inductive argument. Specifically, \( \tilde{H}^{(k+1)} \) is Gaussian because both \( \int_0^t L(t, s) \tilde{H}^{(k)}(s) ds \) and \( \int_0^t J(t, s) dW(s, \omega) \) are Gaussian. Because \( \psi \) is a contraction, we know that \( \tilde{H} \) is the almost sure limit of \( \tilde{H}^{(k)} \), which implies weak convergence. Hence, \( \tilde{H} \) is again Gaussian (because the limit of convergent Gaussian processes is again Gaussian). To elaborate, we may consider the characteristic function of \( \tilde{H}^{(k)}(t) \): \( \Phi_k(s) = e^{is\mu_k - s^2 \sigma_k^2}/2 \) (with \( \mu_k \) and \( \sigma_k^2 \) being the mean and variance of \( \tilde{H}^{(k)} \)), which must converge to the characteristic function of \( \tilde{H} \). Convergence of \( \Phi_k(s) \) at all \( s \) implies the convergence of \( \mu_k \) and \( \sigma_k^2 \), which implies that the characteristic function of \( \tilde{H} \) has the form \( e^{is\mu_\infty - s^2 \sigma_\infty^2}/2 \), which concludes the Gaussian distribution.

**Treating the mean and variance of \( \tilde{H} \).** Taking expectation in (2.26) yields

\[
m_{\tilde{H}}(t) = \int_0^t L(t, s)m_{\tilde{H}}(s) ds + K(t), \quad \text{where} \quad m_{\tilde{H}}(t) = E[\tilde{H}(t)]. \tag{2.90}
\]

It remains to show that the FPE \( x = \Gamma(x) \) has a unique solution, where \( x \in \mathbb{C} \) and the operator

\[
\Gamma(x)(t) = \int_0^t L(t, s)x(s) ds + K(t).
\]
We can do so by showing that $\Gamma : \mathbb{C} \to \mathbb{C}$ is another contraction. Specifically, for $x_1, x_2 \in \mathbb{C},$

$$|\Gamma(x_1)(t) - \Gamma(x_2)(t)| \leq \int_0^t |L(t, s)||x_1(s) - x_2(s)| ds \leq L^\uparrow t \|x_1 - x_2\|_t,$$

where the finite upperbound $L^\uparrow$ is given by (A.8). The rest of the proof is similar.

To treat the variance of $\hat{H}$, consider the SVE (2.26) at $0 \leq s, t \leq T$

$$H(t) - \int_0^t L(t, u)H(u)du = \int_0^t J(t, u)dW(u),$$

$$H(s) - \int_0^s L(s, v)H(v)dv = \int_0^s J(s, v)dW(v).$$

Multiplying the two equations and taking expectation yield that

$$C(t, s) = -\int_0^t \int_0^s L(t, u)h(s, v)C(u, v)dvdu + \int_0^t J(t, u)J(s, u)du$$

$$+ \int_0^t L(t, u)C(u, s)du + \int_0^s h(s, v)C(t, v)dv,$$

where $C(t, s) = \text{Cov}(\hat{H}(t), \hat{H}(s))$, or equivalently, an FPE

$$C = \Theta(C),$$  \hspace{1cm} (2.91)

where $C(\cdot, \cdot) \in \mathbb{C}([0, T]^2)$, and the operator

$$\Theta(C)(t, s) = -\int_0^t \int_0^s L(t, u)h(s, v)C(u, v)dvdu + \int_0^t L(t, u)C(u, s)du$$

$$+ \int_0^s L(s, v)C(t, v)dv + \int_0^t J(t, u)J(s, u)du.$$  \hspace{1cm} (2.92)

Using the norm $\|x\|_{T} = \sup_{0 \leq s, t \leq T} |x(t, s)|$, we next prove that $\Theta$ is a contraction. Specifically, for $x_1, x_2 \in \mathbb{C}([0, T]^2)$, we have

$$|\Theta(x_1)(t, s) - \Theta(x_2)(t, s)| \leq \int_0^t \int_0^s |L(t, u)L(s, v)| \cdot |x_1(u, v) - x_2(u, v)| dvdu$$

$$+ \int_0^t |L(t, u)| \cdot |x_1(u, s) - x_2(u, s)| du$$

$$+ \int_0^s |L(s, v)| \cdot |x_1(t, v) - x_2(t, v)| dv$$

$$\leq \left( (L^\uparrow)^2 ts + L^\uparrow t + L^\uparrow s \right) \|x_1 - x_2\|_{T}.$$
The contraction property is guaranteed if we pick some small enough $\Delta T > 0$ such that
$((L^\uparrow)^2 \Delta T^2 + 2L^\uparrow \Delta T) < 1$. According to the Banach contraction theorem, we have the uniqueness and existence in the small block $[0, \Delta T]^2$. The uniqueness and existence of $C(\cdot, \cdot)$ over the entire region $[0, T] \times [0, T]$ can be proved by recursively dealing with small blocks of the form $[i \Delta T, (i + 1) \Delta T] \times [j \Delta T, (j + 1) \Delta T]$.

**Remark 2.5.2 (Numerical Algorithm for $\sigma_H^2(t)$)** The above proof of the existence and uniqueness of the FPE (2.91) automatically suggests the following recursive algorithm to compute the covariance $C(t, s)$ and variance $\sigma_H^2(t)$. To begin with, we pick an acceptable error target $\epsilon > 0$.

**Algorithm:**

(i) Pick an initial candidate $C^{(0)}(\cdot, \cdot)$;

(ii) In the $k$th iteration, let $C^{(k+1)} = \Theta(C^{(k)})$ with $\Theta$ given in (2.92).

(iii) If $\|C^{(k+1)} - C^{(k)}\|_T < \epsilon$, stop; otherwise, $k = k + 1$ and go back to step (ii).

According to the Banach contraction theorem, this algorithm should converge geometrically fast. When it finally terminates, we set $\sigma_H^2(t) = C(t, t)$, for $0 \leq t \leq T$, which will be used later to devise required control functions $c$ and $\kappa_i$. The algorithm to compute the mean $M_H$ is similar.

**2.5.3 Proof of Proposition 2**

First note that the FPE (A.5) specifies a well-defined function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ such that

$$M_H = \phi(K).$$

(2.93)

See the proof of the uniqueness and existence of the SVE (specifically, see (2.85)–(A.8)) for details. (In fact, it is not hard to see that the function $\phi$ in (2.93) is Liptchitz continuous and linear.)

Let $(\kappa^*, c^*) \equiv (\kappa_1^*, \ldots, \kappa_K^*, c^*)$, with $\kappa_i^*$ and $c^*$ given in (A.10) and (A.9). Let $K^*$ and $M_H^*$ be the corresponding version of (A.4) and the mean of $\tilde{H}$. (We know that $K^*(t) = M_H^*(t) = 0$.) So we have

$$\kappa_i^*(t) = \kappa_i^*(t) - M_H^*(t) = z_{1-\alpha_i}\sigma_{\tilde{H}}(t), \quad 1 \leq i \leq K.$$  

(2.94)

Now consider another solution to $(\tilde{\kappa}, \tilde{c})$ to (2.32), with $(\tilde{\kappa}, \tilde{c}) \equiv (\kappa_1^* + \Delta \kappa_1, \ldots, \kappa_K^* + \Delta \kappa_K, c^* + \Delta c)$. Let $\tilde{K}$ and $\tilde{M}_H$ be the corresponding version of (A.4) and mean of $\tilde{H}$. By (2.32), we have

$$\kappa_i^*(t) + \Delta \kappa_i(t) - \tilde{M}_H(t) = z_{1-\alpha_i}\sigma_{\tilde{H}}(t), \quad 1 \leq i \leq K.$$  

(2.95)
Comparing (2.94) with (2.95), we must have
\[ \Delta \kappa_i(t) = \tilde{M}_H(t) - M^*_H(t) \equiv \Delta \kappa(t) \quad \text{for all } 1 \leq i \leq K. \quad (2.96) \]
Hence, any alternative solution to (2.32) (if any) has the form \( (\kappa_1^* + \Delta \kappa, \ldots, \kappa_K^* + \Delta \kappa, c^* + \Delta c) \). Next, \( M^*_H = \phi(K^*) \) and \( \tilde{M}_H = \phi(\tilde{K}) \) imply that
\[ M^*_H(t) = \int_0^t L(t, s) M^*_H s ds + K^*(t) \quad \text{and} \quad \tilde{M}_H(t) = \int_0^t L(t, s) \tilde{M}_H s ds + \tilde{K}(t), \]
which leads to
\[ \Delta \kappa(t) = \tilde{M}_H(t) - M^*_H(t) = \int_0^t L(t, s) \left( M^*_H s - M^*_H s \right) ds + \left( \tilde{K}(t) - K^*(t) \right), \]
or equivalently \( \Delta \kappa = M^*_H - M^*_H = \phi \left( \tilde{K} - K^* \right) \),
meaning
\[ \Delta \kappa(t) = \int_0^t L(t, s) \Delta \kappa(s) ds + \left( \tilde{K}(t) - K^*(t) \right). \quad (2.97) \]
By (2.96) and (A.4), we have
\[ \tilde{K}(t) - K^*(t) = \frac{\Delta \kappa(t) \sum_{i=1}^K \left( \eta_i(t) - \int_0^t \eta_i(s) e^{\mu_i(s-t)} (\mu_i - h_{F_i}(w_i)) ds \right)}{\eta(t)} - \Delta c(t). \quad (2.98) \]
Finally, combining (2.97) with (2.98), we must have, for any \( \Delta \kappa \),
\[ \Delta c(t) = \Delta \kappa(t) \sum_{i=1}^K \left( \eta_i(t) - \int_0^t \eta_i(s) e^{\mu_i(s-t)} (\mu_i - h_{F_i}(w_i)) ds \right) - \eta(t) \left( \Delta \kappa(t) - \int_0^t L(t, s) \Delta \kappa(s) ds \right) = 0, \]
where the last equality above holds by (A.4). Therefore, we can see that \( c \) is indeed unique, but \( \kappa_i \) is only unique up to adding an arbitrary common function \( \Delta \), which is consistent with our intuition. \( \square \)

### 2.5.4 Proof of Theorem 2.3.2

The FCLT limits in Theorem 2.3.1 implies the FWLLN, that is, we have
\[ (H^n_i, V^n_i) \Rightarrow (w_i^e, w_i^e) \quad \text{in} \quad D^2, \quad \text{for} \quad 1 \leq i \leq K, \quad \text{as} \quad n \to \infty, \]
where $e(t) = 1$. To prove part (i) of Theorem 2.3.2, it is sufficient to show that \( \{V^n_i, n \geq 1\} \) and \( \{H^n_i, n \geq 1\} \) are uniformly integrable (u.i.).

We first prove that the queue length $Q^n_i$ is u.i. To do so, note that $Q^n_i$, which is further bounded by the queue length of an $M_t/GI/\infty$ infinite-server model, having arrival rate $\lambda^n_i(t)$ and service hazard rate $\hat{h}_i(x) = \min\{h_i(x), \mu_i\}$. Denote its length by $X^n_{\infty}(t)$. We have $\tilde{Q}^n(t) \leq_{st} X^n_{\infty}(t)$. Because $X^n_{\infty}(t)$ is a Poisson r.v., the u.i. of $\tilde{X}^n_{\infty}(t)$ is straightforward. Specifically, we have

$$
\sup_n \mathbb{E} \left[ (\tilde{X}^n_{\infty}(t))^2 \right] = \sup_n \left[ \frac{\int_0^t \lambda^n_i(t-x)G_i(x)dx}{n} + \left( \frac{\int_0^t \lambda^n_i(t-x)G_i(x)dx}{n} \right)^2 \right] < \infty, \quad (2.99)
$$

where $G_i$ is the CDF having hazard rate $\hat{h}_i$. See Proposition A.2.2 in [28].

Next, we write the PWT

$$
V^n_i(t) = \sum_{j=0}^{Q^n_i(t)} U_j,
$$

where $U_j$ is the time between the $j^{th}$ and $(j+1)^{th}$ departure times of existing waiting customers at queue $i$. Here a departure includes abandonment and entrance to service. Then

$$
\mathbb{E} \left[ V^n_i(t)^2 \right] = \mathbb{E} \left[ \sum_{j=0}^{Q^n_i(t)} (U_j)^2 + \sum_{j \neq k} U_j U_k \right]
\leq \left( \mathbb{E}[Q^n_i(t)] + 1 \right) \left( \ell_i(1) + 1 \right)^2 + \mathbb{E}[Q^n_i(t)] \ell_i(2) \frac{(\ell_i(1) + 1)^2}{(nm^i \bar{\mu})^2}
$$

where $\bar{\mu} \equiv \min_{1 \leq i \leq K} \mu_i$.

Using the bound in (A.13), we have $\sup_n \mathbb{E} \left[ V^n_i(t)^2 \right] < \infty$, which implies u.i. of $V^n_i$. The u.i. of $H^n_i$ is straightforward because $0 \leq H^n_i \leq T$.

The TPoD for class-$i$ customers

$$
P(V^n_i(t) > w_i) = P(\sqrt{n}(V^n_i(t) - w_i) > 0) = P(\tilde{V}^n_i(t) > 0)
\rightarrow P(\tilde{V}^n_i(t) > 0) = P\left( w_i \left( \tilde{H}(t + w_i) - \kappa_i(t + w_i) \right) > 0 \right)
= P\left( \tilde{H}(t + w_i) > \kappa_i(t + w_i) \right) = P \left( Z > \frac{\kappa_i(t + w_i)}{\sigma_{\tilde{H}}(t + w_i)} \right) = P \left( Z > z_{\alpha_i} \right) = \alpha_i,
$$

where the third equality holds by (A.2).
Proof of Corollary 2.3.2. Because the functions $L(t, s)$ and $J(t, s)$ are now separable in $t$ and $s$, SDE (2.26) becomes

$$
\hat{H}(t) = \frac{1}{R(t)} \int_0^t \tilde{L}(s) \hat{H}(s) ds + \frac{1}{R(t)} \int_0^t \tilde{J}(s) d\mathcal{W}(s) + K(t),
$$

(2.100)

where $R(t)$, $\tilde{L}(t)$ and $\tilde{J}(t)$ are specified in Corollary 1. Multiplying $R(t)$ on both sides and differentiating (2.100) yields

$$
\frac{R'(t) - \tilde{L}(t)}{R(t)} \hat{H}(t) dt + d\hat{H}(t) = \frac{\tilde{J}(t)}{R(t)} d\mathcal{W}(t) + \frac{K'(t)}{R(t)} dt + \frac{K(t) R'(t)}{R(t)} dt.
$$

Multiplying $e^{\int_0^t \frac{R'(t) - \tilde{L}(t)}{R(t)} dt}$ on both sides and integrating from 0 to $t$

$$
e^{\int_0^t \frac{R'(v) - \tilde{L}(v)}{R(v)} dv} \hat{H}(t) = \int_0^t e^{\int_0^u \frac{R'(v) - \tilde{L}(v)}{R(v)} dv} \frac{\tilde{J}(u)}{R(u)} d\mathcal{W}(u) + \int_0^t e^{\int_0^u \frac{R'(v) - \tilde{L}(v)}{R(v)} dv} dK(u) + \int_0^t e^{\int_0^u \frac{R'(v) - \tilde{L}(v)}{R(v)} dv} K(u) \frac{R'(u)}{R(u)} du.
$$

or equivalently,

$$
\hat{H}(t) = \int_0^t e^{-\int_u^t \frac{R'(v) - \tilde{L}(v)}{R(v)} dv} \frac{\tilde{J}(u)}{R(u)} d\mathcal{W}(u)
$$

$$
+ \int_0^t e^{-\int_u^t \frac{R'(v) - \tilde{L}(v)}{R(v)} dv} dK(u) + \int_0^t e^{-\int_u^t \frac{R'(v) - \tilde{L}(v)}{R(v)} dv} K(u) \frac{R'(u)}{R(u)} du.
$$

(2.101)

Note that

$$
e^{-\int_u^t \frac{R'(v) - \tilde{L}(v)}{R(v)} dv} = e^{\log R(u) - \log R(t)} e^{\int_u^t \frac{\tilde{L}(v)}{R(v)} dv} = \frac{R(u)}{R(t)} e^{\int_u^t \frac{\tilde{L}(v)}{R(v)} dv}.
$$

(2.102)

Combining (2.101) and (2.102) yields the solution in (A.15). The variance formula in Corollary 2.3.2 easily follows from the isometry of the Brownian integral. □

2.6 Concluding Remarks

In this chapter, we studied a service differentiation problem for a time-varying queueing system with multiple customer classes. Motivated by call center and health care applications, we measure class-dependent service levels using the so-called TPoD, that is, the probability the waiting time exceeds a delay target. Under a many-server asymptotic framework, we proposed a TV-SRS rule and a TV-DPS policy that can asymptotically achieve TPoD-based performance stabilization, across all customer classes, at any time $t$. Our new TV-DPS rule is both state dependent (based the real-time elapsed cus-
cumer delays) and time dependent (capturing a time variability from the arrival processes). Supple-
menting our limit theorems on asymptotic differentiation and stabilization (Theorems 2.3.1 and 2.3.2),
we also conduct extensive simulation experiments to provide engineering confirmation and practical
insights. In our numerical studies, we used a two-class example to examine the performance of the
proposed policy under a variety of settings. Numerical results show that our proposed solution works
effectively in a wide range of model settings. We also conducted experiments for a five-class example
motived by the CTAS system showing that the algorithm can be easily extended for many classes.
Chapter 3

Modeling and Analysis of Battery Swapping Station for Electric Vehicles

3.1 Introduction

In this chapter, we conduct performance analysis and develop control policy recommendations for the management of continuous battery charge inventory levels of a battery swapping station (BSS). This work is undertaken to provide both a continuous time and continuous charge representation of a BSS that can be used to achieve desired system performance. To that end, we first construct a novel stochastic queueing model representation to capture the continuous inventory levels as a function of wait times in queue. We then develop fluid model approximations to capture the average system information. Then utilizing the stochastic framework developed, we propose both a fluid model method and simulation-based reinforcement learning algorithm to determine charge control policies that will stabilize the quality of battery offered by the BSS in the presence of time-varying demand. This work provides promising first steps for methods of sophisticated, continuous BSS analysis and control.

3.1.1 Background and Motivation of Battery Swapping Stations

According to the United States Environmental Protection Agency, 28% of all of the country’s greenhouse gas emissions in 2016 came from the transportation sector [97]. To combat these high emission levels, many concerned consumers are trading in their traditional combustion vehicles (CV) for the low emission electric vehicles (EV). According to the International Energy Agency, the total number of EVs on the road around the world will grow from 3 million in 2017 to 125 million by 2030. This indicates promising growth in the usage of EVs. However, there are still some challenges facing EV ownership that has deterred a more widespread acceptance.

A time-consuming charging process that takes on the order of hours is unappealing to CV owners
that can refuel in a matter of minutes at a gas station. Furthermore, the existing gas station infrastructure ensures that CV owners can undertake trips that cover long ranges as they have the ability to quickly refuel and continue travelling. EVs are limited in range by the time-consumption investment needed when a recharge is required. Tesla is attempting to combat these issues by developing Supercharger stations in populated urban areas. These stations can provide up to 135 kilowatts (KW) of power and are able to charge a battery to 80% in 45 minutes [89]. However, there are downsides to this quick, high-powered charging scheme as these charging sessions place sharp increases in load on the power grid at random intervals and can lead to battery quality degradation [86]. Even with a more conventional charging session where the EV is plugged in after work during the afternoon and evening, the impacts on the electric grid by EVs can present an additional heavy load at already peak load hours [80]. Energy costs are often much higher in response to these periods of high consumption [2]. However, EV owners might not have an option but to plug in their vehicle after work, thus incurring an undesirably high cost of charging. This creates a high charging maintenance cost of owning an EV that can further discourage ownership.

In response to these issues, a shared-economy alternative has arisen in the form of battery swapping stations (BSS). A BSS stores and charges an inventory of batteries as well as provides a service whereby consumers can swap their low-charge batteries for higher charged ones in a matter of minutes. This can provide the quick refuel experience for EVs that a gas station provides to CVs. This can eliminate a potential consumer’s anxiety about taking long range excursions. Furthermore, the charging burden is transferred from the consumer and placed on the BSS. This creates a centralized hub where smart charging decisions can be made to both reduce the overall charging cost incurred by the BSS while still effectively meeting customer demand, and achieving a more balanced load profile on the electric grid. Instead of paying high charging costs, an EV owner can now simply pay for the exchange when a more charged battery is needed. The BSS is a form of shared-economy in that neither the BSS nor the EV owner maintains possession of an individual battery and instead shares it with the network of users. Thus, there is an overall pool of swappable batteries that exists in circulation to be utilized when needed.

3.1.2 Management of BSS Charge Inventory

When a BSS is constructed, it possesses a fixed number of charging bays that are used to charge the batteries. Because the swapping process is a one-to-one exchange, the number of batteries in inventory remains constant. Upon arrival of a customer, the station begins an automated process whereby an eligible battery from the station will be exchanged directly with the battery in the arrival’s vehicle. The BSS proceeds to charge the newly added battery in addition to its original inventory.

A natural operational question that then arises is how to charge the batteries in inventory effectively throughout the day so as to satisfy customer demand. To develop useful charge control policies, it is
first crucial to accurately model how the charge levels in the BSS are varying throughout the day. The charge inventory is continuously evolving as batteries undergo charging. Several exogenous features also play an important role in determining a BSS’s charge inventory. Time-varying arrival rates to the BSS throughout the day will potentially create periods of high demand, leading to less time to charge the inventory between arrivals and low charge inventory. Arriving customers do not bring in fully depleted batteries and random residual charge levels on incoming batteries will affect the energy requirements needed by the BSS to charge batteries to full. We provide a further discussion and empirical evidence to support these features in §3.2.1.

3.1.3 Contributions

The contributions of this chapter are focused on the inventory charge levels of a BSS and can be separated into two directions: (i) performance analysis and (ii) performance stabilization. Our work is set apart from the literature in that we consider both continuous time models and continuous charge inventory levels.

Performance Analysis

We first provide a novel modeling method that captures key performance functions useful in determining time-varying battery charge levels of BSS. Random customer arrivals to a service system lends itself naturally to a queueing model approach. We propose a novel stochastic system that represents the battery inventory of a BSS as arrivals to a queueing model. We consider batteries entering the queue when a customer arrival occurs. Within the queue, batteries are ordered by charge level with the head-of-line (HOL) battery the highest charged. Upon a battery arrival, we instantaneously depart the HOL battery to mimic the exchange process where an arriving customer would receive the highest charged battery. A common metric to study in queueing models is the waiting time of customers. In this model, waiting of batteries can be translated into representative charge information by the age to charge relationship found in the charging function.

Because the arrival behavior of batteries to the queue is allowed to be time-varying, we rely on the asymptotic approximations to the proposed queue. Fluid models are developed to provide deterministic performance function approximations to characterize the battery charge levels in the system. Numerical studies are conducted to test the proposed results. The models developed are then used in the development of charging policies for the inventory batteries in response to time-varying demand to stabilize the performance functions.

Performance Stabilization

In an effort to achieve stabilized inventory charge levels, we first propose a centralized, continuous charge rate control policy. This policy stipulates a continuously, time-varying rate at which all batteries
in the system are to be charged. With this control method, we next construct two algorithms to find a control policy which achieves stabilization. The first is a fluid based approximation method. Building upon the results from the performance analysis work, we create a fluid model approximation under variable charge rates and mathematically solve for the desired control. In addition, we propose a simulation-based reinforcement learning algorithm that will iteratively update to converge to the desired control as a robust alternative to the fluid method. Numerical studies are conducted to test the proposed results. These control policies and solution methods set promising foundations on which to build increasingly sophisticated BSS control.

3.1.4 Literature Review

Modeling a BSS

While there are few studies focused purely on developing models for performance analysis of a BSS, several studies have been done to optimize the operation of a battery charging station. By necessity these studies have developed a variety of modeling approaches to determine BSS behavior that were subsequently optimized. [23] considers a Monte Carlo simulation solution technique to the optimal inventory planning of a BSS. Inventory charges were considered to be updated on a daily basis. [4] attempts to find an optimal charging schedule over a network of BSSs by solving a non-convex nonlinear optimization problem based on charging cost and energy loss of distribution system. [84] considers the operating of a BSS as a sequence of fixed interval length periods. Batteries charge deterministically according to the number of intervals and demand is satisfied only by full batteries. They solve an optimization problem for optimal number of batteries and charging bays to reduce cost. [83] considers an optimization problem where they perform day-ahead scheduling for preset time slots. They employ robust optimization techniques to manage uncertainty of electric prices and battery demand and allow customers to take less than fully charged batteries. The authors consider the ability of a BSS to resell excess battery energy to the power grid. This feature presents interesting avenues for load-balancing with the electric grid but is not considered in our work. See [81] for additional work that considers partially charged battery swapping. [22] employs Monte Carlo simulation to estimate uncontrolled energy consumption of the BSS using hourly number of EVs for battery swapping, charging start times, travel distance of EVs and charging duration.

The work most closely related to ours is a series of papers by Tan et al. that consider modeling the BSS as a mixed queueing network (MQN), first proposed in [93]. Customers arrive according to Poisson process to a swapping queue and drop off a fully depleted battery in exchange for a fully charged one. Customers unable to find a full battery wait until one becomes available. A fixed number of parking spaces available for the waiting customers is set. The depleted battery joins a depleted battery queue until it undergoes an exponentially distributed charging process, after which it joins a full battery queue. Because the arrival rate of customers is considered to be stationary, their work in [88] focuses
on an optimal charging control in steady state of a Markov decision process to achieve cost reduction. However, in [94], the authors develop a continuous time Markov chain (CTMC) representation and show asymptotic properties of the blocking probability as the scale of the number of batteries grows. Building on this work, [89] proposes a time-varying arrival fluid model to approximate the MQN. However, the fluid model still only considers binary battery charge levels, depleted or fully charged. The assumption is that the charging process length can be approximated with a exponentially distributed charging time with the same mean. Our work also considers a time-varying fluid model, but we propose a novel queueing system that utilizes waiting time information to model a continuous range of battery charge levels, including incoming fluid with residual charge.

Recent work [92] formulates the charging decisions of battery swapping station as a mixed-integer program that minimizes charging costs while meeting demand. They make use of a continuously-valued charge rate level as a decision variable to achieve optimal results. We will propose a similar continuous-charge rate, but our work seeks to find an appropriate continuous-time control. [88] incorporates a quality of service metric that seeks to balance cost-minimization with achieving a low balking rate. Balking occurs due to arrivals to a full system where customers are waiting for a sufficiently charged battery. We seek to directly control the quality of service by controlling the average charge levels on available batteries.

**Fluid Model Approximation**

Approximating stochastic systems with fluid models is widely considered in the literature. Fluid models provide continuous, deterministic approximations to the stochastic system by considering the individual jobs in the system to be atoms of fluid. Deterministic equations can be developed to determine the fluid model dynamics and to give a reasonable approximation of the original stochastic queue dynamics. Kawai et. al develop a fluid approximation of a call center model with time-varying arrivals and after-call work in [51]. [30] develops an iterative-staffing algorithm utilizing fluid models to determine appropriate staffing levels to achieve stable service quality in service systems. Our work follows closely the analysis conducted in [67] and makes extensive use of the so called HOL waiting times and fluid dynamic equations derived therein. The HOL waiting time will be a important model feature in determining battery charge levels.

### 3.1.5 Chapter Organization

The organization of the chapter is as follows. In §3.2 we develop the stochastic queueing model representative of a BSS with basic features and determine performance metrics of interest. In §3.3 we develop the fluid model approximation and specify and ODE for HOL charge and key approximating features. We conduct simulation experimentation to determine the accuracy of the fluid model approximations. In §3.4 we develop both a fluid approach and simulation-based reinforcement learning algorithm to
Figure 3.1: Averaged hourly arrival counts of customers to a BSS in Beijing, China

stabilize charge levels of batteries received by customers. We test both methods through simulation experimentation. In §3.5 we provide concluding remarks.

3.2 Stochastic Model for a BSS

In this section, we motivate and construct a stochastic queueing model with realistic model features. These features will be utilized to determine inventory battery charge levels as a function of the time-varying arrival rate of customers. Using the charge levels, we determine expressions for several key performance indicators.

3.2.1 Empirical Evidence of Realistic Features

This work ultimately considers three realistic features of a BSS that greatly impact charge levels. The features are time-varying arrivals, residual battery charge, and general charging functions. They are discussed in turn with motivation provided for each.

Time-varying arrivals

Key to effectively modeling the charge levels in a BSS is the accurate representation of the arrivals to the system. Arrivals determine when the total charge level in the system jumps. This jump is a reflection of battery charge lost due to the exchange process. If the demand on the system is high, the BSS has less time to charge up the batteries between arriving customers. Thus, capturing periods of high and low demand is crucial to modeling inventory charge levels. We consider the arrivals to be time-varying throughout the day. We are motivated to do so by the hourly arrival counts for a BSS in Beijing, China serving a fleet of taxi vehicles. Figure 3.1 shows that rate at which taxi’s sought a higher charged battery
varied greatly depending on the time-of-day. Peak times included right before lunch and at approximately 8pm. These peak times are offset from the typical peak transportation times of commuters in the morning and in the afternoon. Evidently, taxi drivers are frequenting the BSS to get better charged batteries before their peak operating times in the afternoon. At the end of the day, they pick up a better battery for use the following morning. This presents interesting charging opportunities for the BSS as there is no large demand on the system until mid-morning.

**Residual charge**

Also impacting charging requirements of a BSS is the presence of residual charge in an arrival’s offered battery. When customers seek out an exchange of batteries from the BSS, typically, they are doing so preemptively before the charge on their current battery is fully depleted. If a battery comes into the station with a high residual charge, this can greatly reduce the actual charging requirement of the station. In addition, the battery with the high residual charge will be ready to redistribute to a future arrival quicker than a battery that arrived almost depleted. Figure 3.2 gives a histogram of residual charges of arriving batteries to a BSS in Beijing, China. The average residual charge appears to be about 30% with a reasonable range of incoming charges to be considered from 20-40%. This data is inline with the preemptive behavior of swapping batteries. Taxi drivers will seek out a new battery once their battery has dropped below 50% charged. Not only does this data motivate the consideration of residual charge in system analysis, it suggests that residual charge should be treated as a random quantity associated
Figure 3.3: Charging level of an inventory battery as a function of time in a BSS in Beijing, China with a customer arrival.

**General charging functions**

There is a direct connection between how long a battery has spent in a charging bay and what charge it currently possesses. However, this relationship does not have to be linear. In Figure 3.3, the BSS in Beijing, China has an approximately piecewise linear charging function. When the battery is below 80% charged it receives charge at about 1% per minute. Once the battery reaches the threshold, the charging is scaled back to be about 0.5% per minute. This indicates a functional relationship between age in the system and charge. We will heavily exploit this relationship in the development of later models.

One way to achieve desired system performance is to directly control this charging function. By manipulating the charging function, more charging priority can be given to the lower charged batteries (increasing overall charge) or to the higher charged batteries (increase the best available battery for the next arriving customer) By increasing the rate of charge for lower charged batteries, the overall average system charge can be increased. See §3.1.4 for a discussion of works that attempt to provide optimal charging policies.

### 3.2.2 Base Model Formulation

Motivated by the findings in §3.2.1, we will develop stochastic models considering all three realistic features. Our initial base model focuses first on determining the impact of time-varying arrivals on BSS charge levels and showcasing the effectiveness of the fluid model approach in a finite time interval $[0, T]$. 
To that end, we define a basic stochastic model simulating the behavior of a waiting queue that considers
time-varying arrivals, no residual charge, and a linear charge rate.

There are a fixed number of battery charging slots in a BSS and, as in [70], we will assume the
number of batteries in inventory is equal to the number of charging slots. We let the inventory size be
\( n > 0 \). At a given time \( t \), each battery has a charged battery life of \( L_{(i)}(t) \in [0, 1] \), \( 1 \leq i \leq n \),
where 0 and 1 represent 0% and 100% charge respectively, where \( (i) \) refers to the \( i^{th} \) most charged
battery battery in the system. The BSS begins the day with \( n \) initial batteries (IB). Each IB has an initial
charged life of \( L_{(i)}(0) \in [0, 1] \), \( 1 \leq i \leq n \), where \( (i) \) refers to the \( i^{th} \) most charged IB.

At a given time \( t \), each battery has an age in system of \( W_{(i)}(t) \geq 0 \), \( 1 \leq i \leq n \), where \( i \) refers
to the \( i^{th} \) most charged battery. Batteries charge according to monotonically increasing function \( C(t) \),
with \( C(0) = 0 \) and \( C(T_0) = 1 \), where \( T_0 \) is the time until an battery with 0 battery life becomes fully
charged (i.e., to 100% life). We first study a rate-1 linear charging function \( C(t) = \min\{t, 1\} \). We have
the resulting relation between battery age and battery charge.

\[
L_{(i)}(t) = C(W_{(i)}(t)) = \min\{W_{(i)}(t), 1\} \quad 1 \leq i \leq n.
\]  \hspace{1cm} (3.1)

We let customer arrivals begin at \( t = 0 \) and customers arrive independently according to a non-
homogeneous Poisson process (NHPP) \( N(t) \) with rate \( \lambda(t) \) and cumulative arrival function as
\( \Lambda(t) = \int_0^t \lambda(s)ds \). Each customer brings a fully depleted battery, i.e., residual battery life \( \gamma = 0\% \). Upon arrival
to the system, customers will exchange their empty battery for the best available battery found in the
system. The exchange process time is negligible and will be considered to be 0 as in [108].

Our stochastic model for the BSS is inspired by a traditional customer waiting queue with capacity
\( n \). Batteries arrive to the waiting queue by means of customer arrival, where they will remain until taken
by some future arrival. The amount of time a battery spends waiting in the system before being taken
can be modeled as the battery charging time. The batteries are ordered in system by charge so that the
next customer arrival will always take the highest available charged battery. This can be considered as
ordering the batteries in queue so that the head-of-line (HOL) battery has the highest charge. When
customers arrive, we simultaneously have an arrival of a depleted battery joining the end of queue and
dereparture of the HOL battery from the system, thus keeping total number of batteries constant at \( n \). We
provide a graphical representation in Figure 3.4.

\textbf{Remark 3.2.1 (Difference from conventional service queue models)} In conventional service queues,
each customer, upon arrival, occupies the service resource (measured by the busy time of the server)
for some random time. Besides the waiting time (time a customer spend in queue), the customer makes
nearly immediate impact to the system’s congestion level (i.e., quickly adding new workload to the
system’s existing workload). In our BSS model, rather than staying in the system to occupy the server
for some time, each customer is quickly departed with a “prepaid” service time, leaving the system
some kind of deferred “service deficit”. This, from the customers’ perspective, is quite efficient since
no one has to wait. However, from the service provider’s perspective, the servers have to later work on eliminating that service deficit in preparation of future customers. This observation becomes especially interesting in face of a time-varying demand pattern; careful analysis of the time lags between the arrival peaks and the system’s congestion process can help practitioners better forecast performance and plan capacity.

In addition, conventional service queues consider arrival and departure events as independent. An arrival to the system does not affect the amount of time until the next the customer departure. In contrast, our BSS model has a coupled arrival and departure process for the batteries in system (i.e. an arrival of a battery to the system indicates an immediate departure of another battery).

3.2.3 System Processes

To determine system state as a function of time $t$, we develop expressions for time-varying several stochastic processes. For the rest of this chapter, we will affix an index of $n$ to all stochastic functions to indicate that the process belongs to a system of battery inventory size $n$. Later, we will create a sequence of BSS models with scale ($n$) growing to infinity. This is done to facilitate the proof of convergence of the proposed BSS stochastic model to a deterministic fluid model (proof given in §B.1. All processes are defined on a common probability space. We consider a finite time interval $[0, T]$.

**Customer Arrival Process**

Let $N_n(t)$ be the number of customer arrivals in the interval $[0, t]$ for the $n^{th}$ system. We assume $N_n$ satisfies a FWLLN and FCLT

$$\tilde{N}_n(t) \equiv n^{-1}N_n(t) \Rightarrow \Lambda(t) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty \quad (3.2)$$

$$\hat{N}_n(t) \equiv n^{-1/2}(N_n(t) - n\Lambda(t)) \Rightarrow c_{\lambda}B_{\lambda}(\Lambda(t)) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty \quad (3.3)$$
where \( \Lambda(t) = \int_0^t \lambda(s) \, ds \), \( B_\lambda \) is a standard Brownian motion (BM), and \( \lambda(s) \) and \( c_\lambda > 0 \) measure the time-varying rate of the arrival process and the stochastic variability of the arrival process \( N_n \) asymptotically. Here \( \Rightarrow \) denotes a convergence in distribution. For our work, \( \mathcal{D} \) is the space of cad-lag functions with real-valued right-continuous functions and left limits on the interval \([0, T]\), mainly \( \mathcal{D} \equiv \mathcal{D}([0, T]; \mathbb{R}) \). We assume the arrival rate function is bounded above and positive throughout the interval, \([0, T]\) with \( \lambda^+ = \sup_{t \in [0, T]} \lambda(t) < \infty \) and \( \lambda^- = \inf_{t \in [0, T]} \lambda(t) > 0 \). Every customer arrival corresponds to a battery arrival to queue. By rearranging the definition in (3.2) we achieve the expression

\[
N_n(t) = n^{1/2} \hat{N}_n(t) + n \Lambda(t).
\]  

(3.4)

**Battery Arrival Process**

We introduce separate processes for arriving customers and entering batteries as we will later relax the assumption that every potential customer arrival corresponds to a battery arrival. Let \( E_n(t) \) and \( D_n(t) \) be the total number of batteries that have arrived and departed the \( n^{th} \) system by time \( t \), respectively. Because every customer who arrives enters the system undergoes an instantaneous swapping of batteries, the relation between battery and customer arrivals follows

\[
N_n(t) = E_n(t) = D_n(t) \quad t \geq 0.
\]  

(3.5)

This follows logically as for the number of batteries in system to remain constant we must have arrivals equal to departures with both events corresponding to a customer arrival. Because we assume an instantaneous service time for a customer entering the system, the battery entering process also satisfies a FWLLN

\[
\hat{E}_n(t) \equiv n^{-1} E_n(t) = \hat{N}_n(t) \Rightarrow \Lambda(t) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty,
\]  

(3.6)

by convergence assumed in (3.2). Every customer departure corresponds to a battery exiting the queue.

**Dependency of Interarrivals**

A key relationship to creating these functions is the dependent relationship of customer arrivals in our stochastic model. When the \( i^{th} \) customer enters the system and leaves a battery behind, it is arrived to the end of the waiting queue. We refer to the physical battery left behind by the \( i^{th} \) arrival as \( P_i \). The following \((n-1)\) customer arrivals will arrive batteries that will join the queue behind the \( i^{th} \) customer’s battery, while also removing \((n-1)\) batteries from the head of the line, as a result making \( P_i \) the HOL battery. Therefore, the \((i+n)^{th}\) arrival will always take battery \( P_i \) and the amount of time \( P_i \) spends in the system depends directly on the \( i \) and \((i+n)^{th}\) customer arrival.

We represent the set of customer arrivals that correspond to batteries in the system at time \( t \), as \( \xi(t) \)
which we define as
\[ \xi_n(t) = \{ j | j > (E_n(t) - n)^+ \} \quad t \geq 0. \] (3.7)

At \( t = 0 \), all batteries in the system are IB and we have \( |\xi_n(t)| = 0 \). \( \xi_n(t) \) will continue to increase until all the IB have been taken from the station and we have stabilization of \( |\xi_n(t)| = n \).

**Accrued Battery Life**

The life of the \( i \)th most charged battery in the system at time \( t \) is governed by the most recent \( n \) customers whose exchanged batteries are currently charging in the system. The latest customer arrival corresponds to the newest battery in the system and thus the lowest charge. We let \( L_{j,n}(t) \) be the life of the battery left behind by the \( j \)th customer at time \( t \). We can write \( L_{j,n}(t) \) in terms of the most recent arrivals to the system, as well as any IB that still reside. Note that IB will only remain in the system if \( E_n(t) < n \). We have

\[ L_{(i)}(t) = L_{E_n(t)-n+i}(t)1\{E_n(t) \geq n - i + 1\} + \min\{L_{(E_n(t)+i)}(0) + t, 1\}1\{E_n(t) < n - i + 1\}. \] (3.8)

The first term corresponds to the recently arrived batteries and the second term corresponds to the initial batteries remaining in the system at time \( t \). Initial batteries remaining in system will have an age of \( t \) in addition to the initial life \( L_{(i)}(0) \). In the same way we can write the life of customer \( j \) battery at time \( t \) as a function of the \( j \)th arrival time and we have

\[ L_{j,n}(t) = \begin{cases} \min\{t - \eta_{j,n}, 1\}, & j \in \xi_n(t) \\ 0, & \text{o.w.} \end{cases} , \]

where the \( \eta_{j,n} > 0 \) are the random arrival times of customers to the system. We define \( L_{j,n}(t) \) to be 0 for all \( t \) where the battery is not present in the system.

**Arrival Interval of Current Batteries**

A key result that will be used extensively in later functional developments is the bounding the range of arrival times of batteries that currently exist in the system. To do so, we first define the highest age level (HAL) in the system at time \( t \), \( W_{n}^{\uparrow}(t) \), as the age of the HOL battery (i.e. oldest),

\[ W_{n}^{\uparrow}(t) = (t - \eta_{n-E_n(t)+1})1\{E_n(t) \geq n\} + (L_{(1)}(0) + t) \cdot 1\{E_n(t) < n\} \] (3.9)

Because the battery with the HAL is the oldest battery, its arrival time, \( \eta_{n}^{\uparrow} \), corresponds to a lowerbound on all other arrivals currently in the system, where \( \eta_{n}^{\uparrow}(t) = (t - W_{n}^{\uparrow}(t))^+ \). When \( \eta_{n}(t) = 0 \), the oldest
battery in system is an IB that started the time period in the BSS. The arrival time of all batteries in
the system at time $t$ will correspond to customer arrivals that occur after the HOL arrival time. We can
restrict the possible range of values for the arrival times of batteries in the system at time $t$ as

$$
\eta_{j,n} \in \left[\eta_{n}^\uparrow(t) \right], 
j \in \xi_n(t). 
$$

(3.10)

### 3.2.4 Key Performance Metrics

In this section, we define key performance indicators of the model in an effort to inform BSS managers
as to their ability to satisfy short and long term demand throughout the day.

**Highest Charge Level**

To describe the system’s ability to satisfy short term demand, we wish to capture for time $t$ the highest
charge level (HCL) $L_n^\uparrow(t)$ in the system. This corresponds to the HOL battery life of our model. Because
of the cyclical relationship of arrivals, we know the HOL battery corresponds to a customer who arrived
$n$ customers previous from the total number of arrivals. If less than $n$ number of customers have arrived
at time $t$, then the HOL charge is found in a battery that was initially in the system. We refer back to
(3.8) for the HOL charge.

$$
L_n^\uparrow(t) = L(1)(t) = \min \{t - \eta_{E_n(t) - n + 1}, 1\} \{E_n(t) \geq n\} + \min \{L(E(t) + 1)(0) + t, 1\} \{E_n(t) < n\}. 
$$

(3.11)

When the HCL is less than one, the age and charge of batteries in the system can be used interchangeably.
However, if the HCL is equal to one, the charge level becomes insufficient to determine battery age due
to the minimization in (3.1). The age of a battery is sufficient to determine the charge and we have an
alternative expression for HCL as

$$
L_n^\uparrow(t) = \min \{W_n^\uparrow(t), 1\}. 
$$

(3.12)

**IB Depletion**

To determine at what time the BSS’s initial stock of batteries is depleted from the system (i.e., replaced
by new arrivals), we define $B_{0,n}(t)$ as the total number of initial batteries that still remain in the $n^{th}$
ecosystem at time $t$. This process will be useful in determining when the initial battery ages no longer
affect system processes. The first $n$ customer arrivals will take the $n$ initial batteries first and we have
the remaining initial batteries in system at time $t$ as

$$
B_{0,n}(t) \equiv (n - N_n(t))^+, 
$$

(3.13)
where $(\cdot)^+$ is $\max\{\cdot, 0\}$. We have that $B_{0,n}(t) \in [0,n]$ for all $t \in [0,T]$. We immediately have a FWLLN result for initial battery counting process by continuous mapping theorem applied to (3.2) as

$$\bar{B}_{0,n}(t) \equiv n^{-1}B_{0,n}(t) \Rightarrow (1 - \Lambda(t))^+ \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty \quad (3.14)$$

The time of initial battery depletion, $\Pi_n$, in the $n^{th}$ system is the arrival time of the $n^{th}$ customer (i.e. the $n^{th}$ arrival will depart the last initial battery from the system). Equivalently, this is the time at which the process $B_{0,n}(t)$ first reaches 0.

$$\Pi_n = \inf_{t \geq 0} \{t \mid B_{0,n}(t) = 0\}. \quad (3.15)$$

The interval $[0, \Pi_n]$ can be thought of as the initial warm-up period of the BSS, after which the system is completely determined by the customer arrival process.

If the HCL is less than the current time $t$, the HCL must correspond to a new battery and there must necessarily be no IB remaining in the system. We have the following equivalent statements that indicate the system has only new arrival batteries,

$$\{E_n(t) \geq n\} \iff \{t \geq \Pi_n\} \iff \{B_{0,n}(t) = 0\}. \quad (3.16)$$

**Total Charge Level (TCL)**

To determine how well a BSS can handle long term demand, we capture the charge inventory level of the system at time $t$ (i.e., the sum of all battery charges in the system). We let $Y_n(t)$ be the total charge in the station and is a function of both new and initial battery charges, specifically

$$Y_n(t) = \sum_{i=1}^{n} L(i)(t) \quad (3.17)$$

$$= \sum_{i=1}^{n} L_{E_n(t)-n+i}(t) + \sum_{i=1}^{n\cap n-E_n(t)} \min\{L_{(E_n(t)+i)}(0) + t, 1\} \quad (3.18)$$

$$= \sum_{j=E_n(t)-n+1}^{E_n(t)} L_{j,n}(t) + \sum_{i=1}^{B_{0,n}(t)} \min\{L_{(E_n(t)+i)}(0) + t, 1\} \quad (3.19)$$

$$= \sum_{j=E_n(t-W^+_n(t))}^{E_n(t)} \min\{t - \eta_{j,n}, 1\} + \sum_{i=1}^{B_{0,n}(t)} \min\{L_{(n-B_{0,n}(t)+i)}(0) + t, 1\}, \quad (3.20)$$

where the last equality follows by the fact that the interval $[t - W^+_n(t), t]$ represents the arrival range of new batteries in the system.
Charge Distribution

Let $B_n(t, \alpha)$ be the total number of batteries in the station at time $t$ that have a charge less than $\alpha \cdot 100\%$ and $\alpha < L^+(t)$. $B_n(t, \alpha)$ provides a more detailed description of how the TCL is distributed among the batteries in the system.

$$B_n(t, \alpha) = \sum_{i=1}^{n} 1\{L(i)(t) \leq \alpha\}$$  \hspace{1cm} (3.21)

$$= E_n(t) - E_n(t - \alpha) + \sum_{i=1}^{B_0,n(t)} 1\{L(n-B_0,n(t)+i)(0) + t < \alpha\}$$  \hspace{1cm} (3.22)

$$= E_n(t) - E_n(t - \alpha) + \sum_{i=1}^{B_0,n(t)} 1\{L(n-B_0,n(t)+i)(0) + t < \alpha\},$$  \hspace{1cm} (3.23)

where the last equality follows due to the restriction placed on the arrival time $\eta_{j,n}$ to be greater than $t - \alpha$ in the summation indicator. This restricts the time window for eligible arrivals further. Note that if $\alpha \geq L^+(t)$ we have all battery charges immediately less than $\alpha$ and $B_n(t, \alpha) = n$.

3.3 The BSS Fluid Model

In this section, we define the deterministic fluid model for the BSS queue in an effort to develop tractable functions for average stochastic system behavior. We show in §3.3.5 the effectiveness of our model through computer simulation. Similar to [67], we will characterize our fluid model by a two-parameter density function, parameterized by both time and age of fluid in system. Dynamics of the fluid model are expressed in terms of deterministic, continuous functions and ordinary differential equations (ODE). A graphical depiction of the fluid model approximation technique is show in Figure 3.5.

3.3.1 Fluid Model Dynamics

To fully describe the mechanics of the fluid model, we assume the following structural functions and conditions.

Model Input Parameters

**Arrival and Departure process** The BSS fluid model is represented by a single facility, which mirrors the structure of the waiting room facility in [65]. Incoming fluid flows directly into the entrance of the facility according to a time-dependent arrival rate function $\lambda(t) > 0$. The fluid exits the facility at a rate $\gamma(t) > 0$ and we assume that fluid exits in a FCFS manner, where the oldest fluid exits first. To mimic the arrival and departure processes of the stochastic BSS model, where an arrival to the queue
Figure 3.5: Graphical depiction of the fluid model approximation technique for the BSS queue at a
fixed time $t$

determines an instantaneous HOL departure, we impose the condition that the time-dependent fluid
arival rate function, $\lambda(t)$, is always equal to the time-dependent departure rate, $\gamma(t)$.

$$\lambda(t) = \gamma(t) > 0 \quad t \geq 0.$$  \hfill (3.24)

Because input is equal to output, the total fluid content in the system remains constant for all
$t$. This is consistent with the fixed inventory of $n$ in the stochastic model. We set the total fluid content of the
system to be 1. We restrict our focus to the arrival rate process in rest of this chapter, and use the arrival
rate as representative of both arrival and departures. The total fluid content to enter and leave the system
over the interval $[0, t]$ is given as $\Lambda(t) = \int_0^t \lambda(s) ds$. Because total fluid capacity is 1,
$\lambda(u)$ can be thought of as the rate of fluid in proportion to total fluid capacity entering and leaving the system.

*Initial Conditions* At time 0, the initial fluid content of the model is assumed to follow a density
function, $\tilde{b}(0, x)$, where $x$ is the “age” of initial fluid at time 0. We assume $\tilde{b}(0, \cdot)$ satisfies the following
regularity conditions,

$$\tilde{b}(0, x) > 0 \quad x \in [0, 1], \quad \int_0^1 \tilde{b}(0, x) dx = 1.$$  \hfill (3.25)

The first term enforces the positivity of fluid content for all initial ages and the second term ensures that
the initial fluid content is satisfied to be 1. The initial condition of the system at time 0 can be described
fully by the initial densities of the queue. These densities are model parameters to be specified. To fully
specify the model input parameters, we need only provide the tuple \((\lambda, \bar{b}(0, \cdot))\).

**Key Performance Descriptor**

*Fluid Age Distribution* We let \(\bar{B}(t, y)\) be the total fluid in the system at time \(t\) with age less than \(y\) yielding a representation

\[
B(t, y) = \int_0^y \bar{b}(t, x)dx \quad t \geq 0,
\]

where \(\bar{b}(t, x)\) is the fluid density of age \(x\) at time \(t\). Because total fluid is set to 1, we can also think of \(\bar{B}(t, y)\) as the proportion of total fluid with age less than \(y\). At any given time \(t\), the age distribution function completely determines the makeup of fluid residing in the system. We chose to model age of fluid in the system as it will provide mathematical convenience in later sections. The density of fluid that exists in the system with 0 age at time \(t\) refers to the density of incoming fluid to the system at time \(t\). We treat the arrival rate as the density of incoming 0 age fluid. The crucial relationship to fluid density and arrival process is given as

\[
\bar{b}(t, 0) = \lambda(t) \quad t \geq 0.
\]

**Age to Charge Relationship**

While we will develop the fluid dynamics in terms of fluid age, we need an age to charge conversion in order to effectively approximate the charge levels of the original system. To that end, we translate the age of an atom of fluid into a corresponding charge level through the same relationship as given in (3.1). Thus an atom of fluid with age \(x\) will have a charge of \(\min\{x, 1\}\). It is worth noting that the age and charge of an atom of fluid are equal when the charge is below full capacity. We will discuss more the switch between these two regimes after a discussion of system functions. See §3.3.3 for our treatment of a general charge function.

**3.3.2 Performance Functions**

We define key performance indicator functions to determine the quantities of interest in the system.

**Highest Fluid Charge Level (HFCL)**

Just as in the stochastic model, we are interested in determining what the best available charge in the system is at time \(t\). This charge is carried by the fluid that has resided in the system the longest. Therefore, we are interested in capturing the dynamic behavior of the HOL fluid age, \(w(t)\). We propose that
the HOL fluid age is the solution to the following ODE

\[ \dot{w}(t) = 1 - \frac{\lambda(t)}{\tilde{b}(t, w(t))} \]  

(3.28)

where

\[ \tilde{b}(t, x) = \lambda(t - x)1(t \geq x) + \tilde{b}(0, x - t)1(t < x), \]  

(3.29)

and \( \tilde{b}(t, x) \) can be thought of as the density of fluid at time \( t \) with age \( x \) assuming no fluid departs the system. This ODE was derived following the careful consideration of [65]. Later we will show that this ODE is in fact the limiting equation for the stochastic BSS model. Finally, using the age to charge relationship, we can describe the HFCL, \( \bar{L}^{\uparrow}(t) \), as

\[ \bar{L}^{\uparrow}(t) = \min(\bar{w}(t), 1). \]  

(3.30)

**Overloading and Underloading**

Because the charge on a battery is the truncation of battery age, there exists two regimes of the system as relates to the HOL charge, when \( \bar{L}^{\uparrow}(t) < 1 \) and \( \bar{L}^{\uparrow}(t) = 1 \). We refer to the system as being overloaded when \( \bar{L}^{\uparrow}(t) < 1 \) and underloaded when \( \bar{L}^{\uparrow}(t) = 1 \). When the HFCL is below 1, the system has had an excess of demand leading the BSS to be unable to offer a fully charged battery, thus overloaded system. We develop a method for determining the set of times when the system will switch between stages. Note that in order to be in the overloading case at \( t \), all fluid in the system has to be cycled through at least once in the last 1 time unit. If no full cycle occurred in the last 1 unit, then there must exist some fluid that has resided in the system for longer than 1 time unit, thus yielding a full charge and an underloaded system. We define the function \( \tilde{S}(t) \) where

\[ \tilde{S}(t) = \int_{(t-1)^+}^{t} \lambda(u)du + \int_{0}^{(L^{\uparrow}(t)-t)^+} \tilde{b}(0, u)du. \]  

(3.31)

The first term corresponds to the amount of fluid that has entered the system in the interval \([ (t - 1)^+, t ] \). The second term gives the quantity of initial fluid that is still undergoing charging. The upper limit of the second term comes from the fact that if initial fluid is still in the system it will have a charge greater than or equal to current time because it has been charging for \( t \) units in addition to its initial charge. Only fluid that has an initial age less than \( (L^{\uparrow}(t) - t) \) will still be in the system. We can think of \( \tilde{S}(t) \) as the potential fluid present that is charging at time \( t \) from the arrival of new fluid in the past unit interval and remaining initial fluid. Clearly if \( \tilde{S}(t) > 1 \), more than a unit of fluid has entered the queue and all remaining fluid is young enough to still be charging, thus overloaded. In the overloaded...
regime, \( S(t) < 1 \), less than the full capacity of the system is in a charging state and we are underloaded. The underloaded regime requires careful consideration when translating the age of fluid to charge level. We define the set of potential times of switching between an overloaded and underloaded regime as \( \psi = \{ t | S(t) = 1, t \geq 0 \} \).

**Initial Fluid Depletion**

Similar to the stochastic queue, there is an initial warmup period while the initial battery fluid remains in the system. After this initial fluid has departed the system, the dynamics can be fully determined by the arrival function \( \lambda(t) \). To find this depletion point, we define \( \pi_D \) as the earliest time at which all fluid that started in the system has been removed from the system. This occurs when the total quantity of fluid departing the system is 1, yielding the following expression to be solved for \( \pi_D \)

\[
1 = \int_0^{\pi_D} \lambda(u)du. \tag{3.32}
\]

Note that \( \pi_D = w(\pi_D) \) and because \( \dot{w}(t) = o(1) \) we have the relation \( \{ w(t) < t \} \iff \{ t > \pi_D \} \). Thus, we can simplify the ODE in (3.28) after initial fluid content has been depleted to

\[
\dot{w}(t) = 1 - \frac{\lambda(t)}{\lambda(t - w(t))} \quad t \geq \pi_D. \tag{3.33}
\]

This highlights the singular dependence on the arrival rate after an initial warmup period.

**Total Fluid Charge**

All potential fluid that exists in the system at time \( t \) has an age of bounded above by \( w(t) \). The total charge of fluid in the system is given by

\[
\bar{Y}(t) = \int_0^{w(t)} \min\{ x, 1 \} \hat{b}(t, x)dx \tag{3.34}
\]

\[
= \int_0^{w(t)} \min\{ x, 1 \} \lambda(t - x)dx + \int_{w(t)}^{\infty} \min\{ x, 1 \} \hat{b}(0, x - t)dx. \tag{3.35}
\]

The second equality comes from the breakup of fluid density into the separate treatment of new and initial fluids. This breakup is achieved using the relationship between time of initial fluid depletion and HOL age given in the discussion of (3.33)

**Fluid Distribution**

To determine the total quantity of fluid in the system with a charge above some target level \( \alpha \cdot 100\% \). We need to truncate the valid arrival range from \( [t - w(t), t] \) to \( [t - w(t), t - \alpha] \). By modifying the
upperbound we are ensuring that only fluid that has been in the system for at least $\alpha$ time is considered. The expression is a modification of (3.34)

$$\bar{B}_\alpha(t) = \int_{\alpha \land t}^{w(t) \land t} \min\{x, 1\} \lambda(t - x) dx + \int_{(t - \alpha)^+}^{(t - L(t))^+} \bar{b}(0, x) dx.$$  

### 3.3.3 General Charging Rates

We now extend our model framework to include a general, strictly increasing charging function, $C(w)$, $w \geq 0$ that maps the amount of time a battery has spent charging in system to a corresponding charge level, mainly,

$$C(t - \eta_j) = L_j(t). \tag{3.36}$$

We impose the assumption that $C(\cdot)$ has a well defined inverse, $C^{-1}(\cdot)$. The relaxation of the charging function has no impact on the system mechanics of the base model. The HOL age in the system will evolve in the same manner as before. This relaxation only effects the translation of the age information retrieved from the queue into charge information. Therefore, we can easily modify our existing formulas by using the age-to-life conversion. The performance functions are developed using the HOL age of batteries. For our base case, there existed equality between age and charge when age was less than 1, and we could use age and charge interchangeably. Now we must carefully treat portions of our existing equations that use HOL charge and properly transform it into the age required for the formulation. We give the updated equations for the fluid approximation and provide a simulation example of an exponential charging case.

**Fluid Model**

Recall we defined $w(t)$ as the HOL age of the fluid in the system at time $t$. We have the relation between HOL age and HOL charge as

$$w(t) = C^{-1}(L^\uparrow(t)), \quad L^\uparrow(t) < 1. \tag{3.37}$$

The relationship in 3.37 only holds when the HFCL is less than fully charged. When the HFCL is 1, it is impossible to uniquely determine the HOL age in the system, similar to our discussion in 3.2. After the time at which a battery reaches full charge, $T_0$, the age will map to a full charge of 1. However, we can modify our performance equations by using (3.37). We provide the reevaluated performance indicator functions under the new nonlinear charging policy.
HFCL. We can redefine our HFCL ODE for when $L^\dagger(t) < 1$ as

$$\dot{L}^\dagger(t) = \dot{C}(C^{-1}(L^\dagger(t))) \left( 1 - \frac{\lambda(t)}{b(t, C^{-1}(L^\dagger(t)))} \right), \quad (3.38)$$

where $\dot{C}$ is the derivative of the charge function. This results follows directly from taking the derivative of (3.37) and substituting with (3.28). When $L^\dagger(t) = 1$, the HOL age is needed to calculate the HFCL.

**Total fluid charge.** The total charge in the system is simply modified by the new age-to-charge relationship. To find the appropriate interval of potential fluid in the system, we can convert the HOL age to charge in system. Thus the interval over which arrivals occur that still exist in the system by time $t$ is $[(t - C^{-1}(L^\dagger(t)))^+, t]$

$$\bar{Y}(t) = \int_{(t-C^{-1}(L^\dagger(t)))^+}^{t} C(t-s)\lambda(s)ds + \int_{0}^{(C^{-1}(L^\dagger(t))-t)^+} C(t+x)b(0,x)dx. \quad (3.39)$$

**Fluid distribution** The appropriate interval of fluid to consider in the calculation of fluid at least $\alpha \cdot 100\%$ charged needs to be modified. In the base model, the appropriate interval was $[(t - w(t))^+, (t - \alpha)^+]$. This assumed equality in $\alpha$ charge level and an $\alpha$ length of time charging. Let $T_\alpha$ be the time a battery takes to charge from fully depleted to $\alpha \cdot 100\%$. By the age-to-charge relation, we have $T_\alpha = C^{-1}(\alpha)$. Therefore, we can rewrite the appropriate interval as $[(t - C^{-1}(L^\dagger(t)))^+, (t - C^{-1}(\alpha))^+]$ and have the following expression for charge level distribution

$$\bar{B}_\alpha(t) = \int_{(t-C^{-1}(L^\dagger(t)))^+}^{(t-C^{-1}(\alpha))^+} \lambda(s)ds + \int_{(t-C^{-1}(L^\dagger(t)))^+}^{(t-C^{-1}(\alpha))^+} b(0,x)dx. \quad (3.39)$$

In the following section we will relax the assumption that batteries are brought in fully depleted.

### 3.3.4 Residual Charge Extension

We return to the base model charging function $C(w) = \min\{w, 1\}$ and focus on the relaxation of the assumption that batteries that are brought into the charging station are fully depleted. Let $\gamma_i \in [0, 1]$ be the random residual charge in the $i^{th}$ customer’s battery when it enters the system. We assume $\gamma_i$ are i.i.d. and follow a general CDF and PDF $F(x)$ and $f(x)$ respectively. We also assume an overload interval $[0, T]$.

**Charge Insertion Policy**

With the addition of residual charges, an incoming arrival can potentially have a battery that has a higher charge than the worst battery currently in inventory. In our queueing model representation, the
batteries in the system maintain a sorted order in queue, with the HOL battery being the most charged and the least charged the end-of-line. To maintain this order in the face of residual charge, we add the incoming battery to its appropriate place in line to maintain sorted charge order. We refer to this as the charge insertion policy (CIP) for arriving batteries. With this insertion policy, we can continue to depart the HOL battery from the system upon an arrival. CIP ensures that batteries are properly ordered in queue but disrupts the assumed first-come-first-served policy of the queue. The HOL battery does not necessarily have to be the battery that has resided in the system for the longest amount of time. A later arrival with a sufficiently high residual charge can be taken before an earlier arrival with a low charge.

**Arrival Interval for Batteries in System**

Because the total charge of the HOL battery is a combination of residual charge and age in the system, the interval $[t - L^\uparrow_n(t), t]$ will be larger than the interval spanning from the arrival of the HOL customer to $t$. By including all arrivals in this larger interval, we are potentially including arrivals that occurred before the HOL battery arrived to the system. This is acceptable as our assignment policy departs the battery with the greatest combined charge (residual charge and charge received in queue) not necessarily the oldest battery in the system (i.e. not FCFS). Only the arrival on that interval that corresponds to batteries with an age less than $L^\uparrow(t)$ at time $t$ will be counted. We argue that for those arrivals occurring after the HOL arrival time the range of acceptable indicator holds as any battery that violates the condition must have exited the system or else violate the definition of $L^\uparrow(t)$ because that battery will have been compared to the HOL battery.

However, the arrivals occurring before the HOL battery warrant a careful treatment. An arrival before the HOL battery could potentially have an age less than $\alpha$ if the residual charge is small enough. However, we do not have the same violation of the definition of $L^\uparrow(t)$ to ensure that indicator holds as these batteries arrive before the HOL battery is in the system can be potentially depart the system before the HOL battery arrives (i.e. be the HOL battery before the current HOL battery enters the system and departs). We argue that this case will not happen by utilizing an upperbound on the rate of change on $L^\uparrow(t)$. Because we assume that no battery enters the queue with a residual charge greater than the HOL charge, the HOL charge will evolve at charge rate 1 or decrease by a non-negative charge difference when the HOL battery departs the system, thus we have the upperbound

$$L^\uparrow_n(t + \delta) \leq L^\uparrow_n(t) + \delta. \quad (3.40)$$

We prove by contradiction. Assume that a customer arrives at some time $\eta$ with residual charge $\gamma$ before the HOL customer at time $t$ having arrival time $\eta^*$ and residual charge $\gamma^*$, such that the first customer would have a total charge at time $t$ less than the HOL charge, $L^\uparrow(t)$,

$$(t - \eta) + \gamma < (t - \eta^*) + \gamma^* = L^\uparrow_n(t) \quad \eta < \eta^*. \quad (3.41)$$
Also assume that at some time \( s \), where \( \eta < s < \eta^* \) the first customer will depart the system. By our routing rule, this implies that \( L_n^\uparrow(s-) = \gamma + (s-\eta) \). From the HOL evolution equation (3.40), we have the upperbound for the HOL charge at time evolved from \( s \) to \( t \) as

\[
\hat{L}_n^\uparrow(t) \leq L_n^\uparrow(s-) + (t-s) = \gamma + t - \eta < L_n^\uparrow(t).
\]

(3.42)

We have the upperbound of HOL charge at time \( t \) as less than the HOL charge at time \( t \) which is a contradiction. Thus, a battery that arrives to the system in the interval \([t - L^\uparrow(t), t - \eta^*]\) that would potentially have a charge less than HOL charge at time \( t \), must remain in the system. □

**Fluid Model**

To mimic the CIP of the stochastic queue, we distribute incoming fluid to the charging queue by the residual charge distribution as depicted in Figure 3.6. For the initial analysis, we enforce the constraint on the residual charge distribution that the highest possible charge to enter the system must be lower than the lowest HFCL achieved by the system during the time horizon of interest. This constraint is enforced to ensure that fluid entering the system is not placed ahead of the HFCL. This event would imply that arriving fluid is instantaneously departed from the system.

**HOL Age.** With a distribution of initial charge of fluid coming into the system, we can think of the total incoming rate of fluid as being split into individual rates for each charge according to the residual charge distribution. The total amount of fluid leaving the system at time \( t \) remains \( \lambda(t) \). The amount of fluid joining the queue at time \( t \) with age \( x \) is defined as

\[
\lambda_e(t,x) = \lambda(t)f(x) \quad t \geq 0, \quad 0 \leq a \leq x \leq b < \inf_{t \geq 0}\{L_n^\uparrow(t)\},
\]

(3.43)

where \( f(x) \) is the density function of the residual charge of fluid and \( a \) and \( b \) represent the lower and upper bound on distribution of incoming fluid, respectively. We impose the restriction the entering fluid will never have an residual charge greater than the HOL charge. We modify our two-parameter density function assuming no fluid departure in (3.29) as follows

\[
\tilde{b}(t,x) = \int_0^{t\wedge x} \lambda(t-s)f(x-s)ds + \tilde{b}(0,x-t)1(x>t).
\]

(3.44)

The HOL age ODE remains unaffected and is given as

\[
\dot{w}(t) = 1 - \frac{\lambda(t)}{\tilde{b}(t,w(t))}.
\]

(3.45)
Figure 3.6: Graphical depiction of how the incoming fluid is distributed to the fluid queue by the density function of residual charge under CIP

Total Fluid Charge: To accurately capture fluid charge we need to incorporate the residual charge on incoming fluid,

\[
\bar{Y}(t) = \int_{t-L^+(t)}^{t} \int_{0}^{L^+(t)-t+s} \lambda(s)(x + t - s)f(x)dxds + \int_{0}^{(L^+(t)-t)^+} (t + x)\bar{b}(0, x)dx. \tag{3.46}
\]

Charge Distribution: We can describe the queue age distribution process as using the two parameter process \(B(t, y)\). Note that if \(L^+(t) < y\) then \(B(t, y) = 1\). If \(L^+(t) \leq t\) we have

\[
\bar{B}(t, y) = \int_{0}^{y} \bar{b}(t, x)dx
\]

\[
= \int_{0}^{y} \int_{0}^{t\wedge x} \lambda(t - s)f(x - s)dsdx + \int_{0}^{(y-t)^+} \bar{b}(0, x)dx \tag{3.47}
\]

\[
= \int_{0}^{t\wedge y} \lambda(t - s)F(y - s)ds + \int_{0}^{(y-t)^+} \bar{b}(0, x)dx. \tag{3.48}
\]
Residual Charge with Balking

To relax the assumption that the highest charged battery in system is always less than the HOL charge, we now assume that customers entering the system finding a HOL battery charge lower than the charge with which they enter will decline the exchange and exit the system. In this event, an old battery does not enter the system (i.e. balking). We let \( K_n(t) \) be the number of customers who have declined to exchange batteries by time \( t \) in the \( n^{th} \) system where

\[
K_n(t) = \sum_{i=1}^{N_n(t)} 1\{\gamma_i > L^\uparrow(\eta_i)\} \quad t \geq 0. \quad (3.50)
\]

Because not all customers who attempt to enter the system exchange a battery with the BSS, we have the relation \( N_n(t) = E_n(t) + K_n(t) \). As a result, the number of customers to actually enter the system can be written

\[
E_n(t) = \sum_{i=1}^{N_n(t)} 1\{\gamma_i \leq L^\uparrow(\eta_i)\} \quad (3.51)
\]

\[
= n \int_0^t \int_0^1 \mathbb{1}\{x \leq F(L^\uparrow_n(s))\}d\bar{U}_n(\bar{N}_n(s), x). \quad (3.52)
\]

See Figure 3.7 for a graphical representation of customer arrivals who balk, enter service, and still remain in system by time \( t \).

We now have the charge distribution process given as

\[
B_n(t, y) = \sum_{i=E_n(t-y)}^{E_n(t)} 1\{t - \eta_i + \gamma_i < y\} + B_n(0, (y - t)^+) \cdot \mathbb{1}\{y \geq t\} \quad (3.53)
\]

\[
= n \int_{t-y}^t \int_0^1 \mathbb{1}\{x \leq F(y + s - t)\}d\bar{U}_n(\bar{E}_n(s), x) + B_n(0, (y - t)^+) \cdot \mathbb{1}\{y \geq t\}. \quad (3.54)
\]

Fluid Model: We can describe the balking rate of fluid to the system as

\[
\lambda_b(t) = \lambda(t)F(L^\uparrow(t)) \quad t \geq 0. \quad (3.55)
\]

We modify our two-parameter density function assuming no fluid departure in (3.29) as follows

\[
\tilde{b}(t, x) = \int_0^{t \wedge x} \lambda(t-s)f(x-s)\mathbb{1}\{x-s \leq L^\uparrow(t-s)\}ds + \tilde{b}(0, x-t)\mathbb{1}\{x > t\} \quad (3.56)
\]

\[
= \int_{(t-x)^+}^t \lambda(u)f(u-(t-x))\mathbb{1}\{u-(t-x) < L^\uparrow(u)\}du + \tilde{b}(0, x-t)\mathbb{1}\{x > t\}. \quad (3.57)
\]
The arrival rate of fluid into the system is now dependent on HOL charge and the HOL charge ODE becomes

$$\dot{\bar{L}}^+(t) = 1 - \frac{\lambda(t)F(\bar{L}^+(t))}{\bar{b}(t, \bar{L}^+(t))},$$

where

$$\bar{b}(t, \bar{L}^+(t)) = \int_{(t-\bar{L}^+(t)+t)}^t \lambda(u)f(u - (t - \bar{L}^+(t)))1\{L^+(t) - L^+(u) < t - u\}du$$

$$= \int_{(t-\bar{L}^+(t)+t)}^t \lambda(u)f(u - (t - \bar{L}^+(t)))du + \bar{b}(0, \bar{L}^+(t) - t)1(\bar{L}^+(t) > t),$$

where the second equality follows from the upperbound on change of $\bar{L}^+(t)$ arising from the convergence.
3.3.5 Engineering Confirmation via Simulations

To test the effectiveness of the fluid model approximations developed in §3.3-3.3.4, we developed an instance of a battery swapping system and conducted discrete event simulation to ascertain system behavior. We chose model parameters to ensure the system would be pushed into periods of overloading and underloading. We set \( n = 15 \) and the arrival process to be a NHPP with \( \lambda(t) = 25 + 15 \sin(t) \). A sinusoidal arrival rate function is used to adequately represent a time-varying system. We conducted 500 independent sample paths of BSS operation. Because we are interested in transient behavior, a time horizon of 24 hours was considered. All performance indicators were captured at a time step of \( dt = 0.001 \) to give an approximate view of continuous results. The results were averaged across the simulations to achieve the time-dependent expected behavior of the key performance indicators.

Linear Charge

We first test the base model fluid equations given in §3.3 under the base charging function \( C(w) = \min\{w, 1\} \). Figure (3.8) plots the simulated key performance indicators (blue line) against the fluid model approximation (red line). The first subplot gives the arrival function to emphasize the time-varying nature of the arrivals. The second subplot gives the HAL. The third plot shows the expected HCL as well as the HFCL \( \Lambda^u(t) \). As evidenced by the plot, the fluid model approximates the mean behavior of the system even with a relatively small number of batteries. The fourth subplot gives the total battery life of the system \( Y_n(t) \). This is sum of all total charge present at time \( t \). In the fifth subplot, we give the proportion of batteries that are \( 80\% \) charged. We see that even for small \( n \) the fluid model performance functions match the simulation results. We test the fluid approximation’s robustness to scale in Figure 3.9 and Figure 3.10 where the number of inventory batteries is 5 and 50 respectively. As the number of batteries decrease, the approximation is less able to capture the extreme behavior of the system.

Exponential Charge

We modify the example given in §3.3.5 by changing the charging function to be \( C(w) = 1 - e^{-2w} \). Note that with this charging function, batteries will never reach full charge in finite time. However, we can still capture the proportion of batteries at \( \alpha \) charge level. We conduct another Monte Carlo simulation and compare our analytic formulas to Monte Carlo estimators in Figure 3.11.
Figure 3.8: Simulation comparison (blue line) of performance metrics to the fluid model performance functions (red line) with $n = 15$, $\lambda(t) = 25 + 15 \sin(t)$, and charging function $C(w) = \min\{w, 1\}$.

**Residual Charge**

Batteries entering the system were given a charge according to the PDF $f(x) = 7.5 - 750(x - .01)^2$ for $0 \leq x \leq 0.2$ which is a parabolic density function with finite support on the interval $[0, 0.2]$. This density function was chosen to provide mathematical tractability in calculating performance indicators. Batteries enter the system with a charge ranging from depleted to 20%. We set the charging function to be the linear rate-1 with $C(w) = \min\{w, 1\}$. Figure 3.12 shows that under a random residual charge, the fluid approximation is still valid. The density function of residual charge is given in Figure 3.13.
Figure 3.9: Simulation comparison (blue line) of performance metrics to the fluid model performance functions (red line) with $n = 5$, $\lambda(t) = 8.3 + 5 \sin(t)$, and charging function $C(w) = \min\{w, 1\}$
Figure 3.10: Simulation comparison (blue line) of performance metrics to the fluid model performance functions (red line) with $n = 50$, $\lambda(t) = 83.3 + 50 \sin(t)$, and charging function $C(w) = \min\{w, 1\}$
3.3.6 FWLLN for BSS under General Charging Function

To provide mathematical support for the use of a fluid model approximation, we present a joint convergence result for all system processes in Theorem 3.3.1 under mild initial charge assumptions and no residual battery charge. We allow for a general charge function that is also time-dependent. This will facilitate a charge control as discussed in the following section. The resulting system processes are shown to converge to the deterministic fluid formulas discussed in the previous chapter as the number of batteries in the system grows to be large.

**Theorem 3.3.1 (FWLLN for the BSS base model with zero initial charge)** Consider an interval \([0, T]\) where \(L_{(i)}(0) = 0, i = 1, \ldots, n\) and batteries undergo charging according to a general, continu-
Figure 3.12: Simulation comparison (blue line) of performance metrics to the fluid model performance functions (red line) with $n = 15$, $\lambda(t) = 25 + 15 \sin(t)$, charging function $C(w) = \min\{w, 1\}$, and residual charge density function $f(x) = 7.5 - 750(x - .01)^2$ for $0 \leq x \leq 0.2$

Ours time-dependent charging function $C(w, t)$. If all other assumptions in §3.2.2 hold, then, as $n \to \infty$

$$(\bar{N}_n, W_n, L_n^\uparrow(t), \bar{B}_{0,n}, \bar{V}_n, \bar{B}_n) \Rightarrow (\Lambda, w, L^\uparrow(t), B_0, Y, B) \quad \text{in} \quad \mathcal{D}^5 \times \mathcal{D}_T, \quad (3.61)$$
where \( D_D \equiv D([0, T], D([0, T], \mathbb{R})) \), \( \Lambda(t) = \int_0^t \lambda(s)ds \), and \( w, L^\uparrow(t) \) \( B_0, Y, \) and \( B \) are given by

\[
\dot{w}(t) = 1\{\Lambda(t) < 1\} + \left(1 - \frac{\lambda(t)}{\lambda(t - w(t))}\right)1\{\Lambda(t) > 1\}, \tag{3.62}
\]

\[
L^\uparrow(t) = C(w(t), t), \tag{3.63}
\]

\[
B_0(t) = (1 - \Lambda(t))^+, \tag{3.64}
\]

\[
Y(t) = \int_{t-w(t)}^t \min\{t - s, 1\} \lambda(s)ds + t \left(1 - \int_0^t \lambda(s)ds\right)^+, \tag{3.65}
\]

\[
B(t, y) = \int_{t-(w(t)\wedge y)}^t \lambda(s)ds + \left(1 - \int_0^t \lambda(s)ds\right)^+1\{t < y\}, \tag{3.66}
\]

for \( t \geq 0 \).

This proof follows a standard tightness argument. We give the detailed proof in Appendix B.1.

### 3.4 Achieving Time-Stable Service Quality

In this section, we focus our efforts on effectively managing the BSS charge inventory to achieve time-stable service quality. We give an overview of our control approach and solution techniques in §3.4.1. §3.4.2 and §3.4.3 give the details of a fluid model and simulation-based reinforcement learning algorithm to solve for desired charge control, respectively. §3.4.4 provides numerical experimentation to validate both solution algorithms.
3.4.1 Service Quality Control of a BSS

Stabilization of Expected HCL

To meet a service quality standard, we need to focus our efforts on stabilizing the expected HCL level. In our model, the HCL represents the customer’s sole interaction with the BSS. By stabilizing, the HCL we can ensure that customers will receive the same battery quality no matter what time of day they utilize the BSS service. Due to the stochastic nature of customer arrivals, the HCL is random throughout the day. To ensure customer satisfaction, our goal is to determine a rate control policy that will stabilize the best available battery charge to a desired target level, $L^*$, for a finite time horizon $[0, T]$. Furthermore, pricing of battery swapping is based off of a differential between offered battery and received battery. By ensuring a stabilized HCL, a BSS is, in part, stabilizing the revenue. Furthermore, by charging to stabilize HCL, the system is operating similar to a just-in-time system, whereby the finished products are produced just as demand occurs for them. In a similar manner, by stabilizing the expected HCL, we are ensuring that batteries will be charged just enough so that an arriving customer on average receives the battery just as it is charged to the performance target. This method provides a minimum on the battery charge inventory needed to satisfy demand. This allows managers to charge batteries at potentially slower rates while still meeting the quality charge level. This is beneficial to the long-term health of the charging inventory as high charging rates can cause lasting damage to the physical batteries.

Charge Rate Control

To control the expected HCL, we propose a time-varying continuous charge rate control, similar to Tan et. al. [94]. We choose to centralize our charging decision, that is all batteries in the system will charge simultaneously at the same rate. We assume that at any given time $t$, all batteries in the BSS will experience charging at rate $\mu(t) > 0$ if not fully charged. The amount of charge received by a battery at time $t$ is a function of the time spent in system $w$. We can determine the potential added charge to the battery as

$$\Delta L = \int_{t-w}^{t} \mu(s)ds. \quad (3.67)$$

We propose this charging policy for both ease of implementability in practice (a single, time-dependent control policy) and mathematical tractability. In addition, through simulation experimentation, we also see that other key performance indicators discussed in §3.2.4 (TCL and charge distribution) are effectively stabilized as well, shown in 3.4.4. Achieving stabilization across all three performance indicators is engineering support for the benefit of both our proposed control of HCL and the rate control method. For example, while our model allows for customers to take batteries that are less than fully charged, a potentially useful metric is the average proportion of batteries that are fully charged. If the quality of
service indicator was modified to be proportion of inventory that has finished charging, we are still able to achieve average behavior stabilization through our control of the HCL.

Solution Techniques

To find the desired control function, we propose two methods that each have appeal in different areas. First, we develop a fluid model approximation method. Utilizing the fluid models developed in §3.3, we create a fluid model under continuous rate control and determine an analytical expression for the HFCL. Through careful treatment of this expression, we create a charging policy recommendation for the original stochastic system. However, the effectiveness of the fluid-based policy begins to deteriorate if the scale of the system is small (i.e. only a few batteries in the station). Furthermore, the mathematical basis for a rate-controlled model with residual charge has not yet been established. See Appendix B for more detail.

To augment the fluid method, we take inspiration from the iterative staffing algorithm developed in [29] to develop an algorithm utilizing Monte Carlo simulation. We propose a solution technique that iteratively updates a control policy until the desired model performance is achieved. We refer to our method as the simulation-based reinforcement learning (SBRL) method. The SBRL method is a more robust solution technique as it can be used to find a charge policy for any scale system and model complexity, provided the system can be accurately simulated. However, this technique does not yield a control function, but by necessity of computer simulation, yields a control table. Furthermore, the computational effort required for this method is an increasing function of model size and complexity. In the following section, we will give the development and full algorithms for each solution method.

3.4.2 Fluid Model Algorithm for Performance Stabilization of a BSS with No Residual Charge

For this algorithm, we will use the base model framework developed in §3.2.2. Customers arrive according to a NHPP with rate $\lambda(t)$ with batteries that are fully depleted. However, instead of a constant rate-one charging function, we assume a general non-negative, time-dependent charging function $\mu(t)$. The age-to-charge relationship in now time-dependent due to the varying $\mu$ and is given as

$$L_{(i)}(t) = \int_{t-W_{(i)}}^{t} \mu(s)ds.$$  \hspace{1cm} (3.68)

Remark 3.4.1 The resulting random charge level of the battery is a non-linear transformation of random battery age. However, the distribution of the battery age is unaffected by the charge function chosen. For the no residual charge case, the BSS under rate control achieves the same battery age distribution as the base model in §3.2.2.
Development and Proof of Fluid Model Charge Function

Exploiting the fact that the age distribution (and therefore HOL age) is unchanged by choice of charge-rate function, we can immediately give the ODE for head-of-line fluid age under rate control as

\[ \dot{w}(t) = 1 - \frac{\lambda(t)}{\lambda(t - w(t))} \quad t > t_D, \quad (3.69) \]

where \(t_D\) is the time of depletion of initial fluid content. By solving this ODE and the age-to-charge relationship given in (3.68), we can write the HFCL as

\[ \bar{L}^{↑}(t) = \int_{t-w(t)}^{t} \mu(s) ds. \quad (3.70) \]

This following proposition gives the fluid charge rate function \(\mu^*\) that will achieve HFCL stabilization.

**Proposition 3 (Fluid Charge Rate Function)** For a BSS system with zero initial charge, charging according to \(\mu^*\) will achieve \(\bar{L}^{↑}(t) = L^*, \quad t_D \leq t \leq T.\)

\[ \mu^*(t) = \begin{cases} \mu_0 + \frac{\mu_D - \mu_0}{t_D} t, & 0 \leq t < t_D \\ \mu(t - w(t)) \frac{\lambda(t)}{\lambda(t - w(t))}, & t_D \leq t \leq T \end{cases} \quad (3.71) \]

where \(\mu_0 = \frac{2L^*}{t_D(1 + \frac{\lambda(t_D)}{\lambda(0)})}\) and \(\mu_D = \mu_0 \frac{\lambda(t_D)}{\lambda(0)}\).

Proof: We first utilize the ODE for HOL age to solve for a control function that will yield a zero derivative, thus ensuring a constant HFCL.

\[ \dot{L}^{↑}(t) = 0 = \frac{d}{dt} \left( \int_{t-w(t)}^{t} \mu(s) ds \right), \quad (3.72) \]

\[ = \mu(t) - \mu(t - w(t))(1 - \dot{w}(t)), \quad (3.73) \]

\[ = \mu(t) - \mu(t - w(t)) \frac{\lambda(t)}{\lambda(t - w(t))}, \quad (3.74) \]

\[ \mu(t) = \mu(t - w(t)) \frac{\lambda(t)}{\lambda(t - w(t))}, \quad t > t_D. \quad (3.75) \]

For time after initial battery depletion, the charge rate function is dependent on the charge rate selected when the current oldest battery first entered the system \((t - w(t)).\) It remains to ensure that \(\dot{L}^{↑}(t_D) = L^*\) and specify the charge rate for the interval \([0, t_D].\) We impose a simple linear change in rate from \([0, t_D]\) where \(\mu(0) = \mu_0\) and \(\mu(t_D) = \mu_D.\) We can solve the following set of equations for \(\mu_0\) and \(\mu_D\) to ensure
continuity of $\mu$:

$$\int_0^{t_D} \mu_0 + \frac{\mu_D - \mu_0}{t_D} s ds = L^*, \quad \text{(target at } t_D)$$

$$\mu(t_D) = \mu(0) \frac{\lambda(t_D)}{\lambda(0)}. \quad \text{(continuity at } t_D)$$

**Remark 3.4.2** $\mu^*$ is a unique function with respect to the selection of charge rate function on the interval $[0, t_D]$. In addition, if the assumption of zero initial charge is relaxed, the relationship given in (3.75) will still provide a suitable charge function given the initial interval charge rate is chosen so that $\bar{L}^*(t_D) = L^*$.

We give the complete algorithm for determining an appropriate charge function using the fluid model approximation in Algorithm 1. We let $\mu^*_F$ be output of this algorithm.

**Algorithm 1**

A Fluid Approximation Algorithm to Determine $\mu^*_F$ for Stabilizing HCL to Constant $L^*$

1: Set $t_D \leftarrow 1 = \int_0^{t_D} \lambda(s) ds$ for $t_D$
2: Set $\mu_0 \leftarrow \frac{2L^*}{t_D(1 + \frac{\lambda(t_D)}{\lambda(0)})}, \mu_D \leftarrow \mu_0 \frac{\lambda(t_D)}{\lambda(0)}$
3: for $t \in [0, T]$
4: if $t \leq t_D$ then
5: Set $\mu^*_F(t) \leftarrow \mu_0 + \frac{\mu_D - \mu_0}{t_D} t$
6: else
7: Set $\mu^*_F(t) \leftarrow \mu^*_F(t - w(t)) \frac{\lambda(t)}{\lambda(t - w(t))}$
8: end if
9: end for

**Numerical Experimentation of Algorithm 1**

**Base Case** We first study a BSS with 50 available batteries and zero residual charge on incoming batteries. We arrive customers to the system according to a NHPP with rate $\lambda(t) = 83.3 + 50 \sin(t)$. To clearly visualize the performance stabilization, a target HCL of 80% was chosen. $\mu^*_F$ was calculated according to Algorithm 1 and multiple sample paths of a BSS were generated under the specified charge rate. The resulting key performance indicators are presented in Figure 3.14 and we see excellent stabilization. We see apparent performance stabilization in other performance metrics, such as total charge inventory and proportion charged to $\alpha\%$. In the bottom two panels of Figure 3.19 we see both TCL and proportion of inventory with greater than 50% charge are also stabilized.

**Small System Scale** To test the robustness of the fluid algorithm to small system scale, we modify the parameters of the previous example so the number of batteries in inventory is 2 and the arrival rate is $\lambda(t) = 3.33 + 2 \sin(t)$. Due to the smaller system size, the variance of the HCL is larger and, as a result,
Figure 3.14: Simulation Results of Fluid Algorithm Performance Stabilization of HCL, TCL, and Proportion Charged at Least 50%: Base Case ($\lambda(t) = 83.3 + 50 \sin(t)$ and No Residual Charge)

A larger sample size is needed to achieve the same approximation precision as the previous example. However, in an effort to reduce computation requirements due to large number of sample runs, $\epsilon$ was instead set to be 0.02 to allow for more stochastic volatility. Figure 3.15 shows that Algorithm 1 is not able to achieve the target expected HCL of 0.8. The fluid method appears to underestimate the charge rate required throughout the time horizon. Interestingly, $\mu^*_F$ still achieves apparent stabilization in all three performance metrics.

**Limitations of Algorithm 1** The fluid based algorithm, while effective in some model settings, has several limitations.

1. **Small System Scale** Because the algorithm is based off of a fluid limiting model, the output of the algorithm decreases in effectiveness with system scale, as shown in Figure 3.15.

2. **Model Complexity** The current mathematical results only support the development of a charge rate function for a BSS under the assumption of no residual charge. We propose necessary expressions
Figure 3.15: Simulation Results of Fluid Algorithm Performance Stabilization of HCL, TCL, and Proportion 50% charged: Small Scale Case ($\lambda(t) = 8.3 + 5 \sin(t)$) and No Residual Charge

for the residual charge case in Appendix B but are unable to provide an implementable Algorithm to solve for $\mu_{T^*}$. Furthermore, any additional model features would require a separate proof to determine appropriate solutions.

3. **High QoS** As the target charge approaches the upper bound of 1 (fully charged), the algorithm will decrease in ability to stabilize the expected HCL to the desired quality of service. In the stochastic system, the error from target value will be bounded above by a difference from 1. As the QoS becomes higher, the error will become strictly negative (less than fully charged) and the resulting average HCL will be strictly less than the target. The fluid approximation assumes a scale large enough so that there is no negative error and, as a result, Algorithm 1 underestimates the charge required to achieve a high QoS.

Motivated to augment Algorithm 1 in the cases where the analytical form fails to perform well, we attempt to build a more practical and robust approach that can also achieve desired performance.
Figure 3.16: A Flowchart for SBRL Algorithm
3.4.3 A SBRL Algorithm for Performance Stabilization

This method utilizes Monte Carlo simulation to estimate mean charge levels across a finite time horizon under candidate charge functions. Using the simulation output, the candidate charge function is modified to better achieve desired service quality. This reinforcement learning process repeats, further refining the charge function until it achieves desired performance stabilization within a tolerance threshold. Figure 3.16 gives a flowchart overview of this reinforcement learning technique.

For the computer simulation we must be able to generate sample paths of the various stochastic processes of a BSS. For the base case, we assume that we can generate arrival times following a non-homogeneous Poisson arrival process characterized by the arrival rate function $\lambda(t), \ t > 0$. Arrival times are generated over a finite time-horizon $[0, T]$. For the residual charge case, we assume the ability to generate IID residual charges according to PDF $f(x)$. We first give the scheme for the proposed algorithm in Algorithm 2 and then discuss its reasoning and development of the algorithm. We let the output of this algorithm be $\mu_{SBRL}^*$.

Algorithm 2 A SBRL Algorithm for Determining $\mu_{SBRL}^*$ for Stabilizing HCL to Constant $L^*$

1: **Initialization:** Set $i = 0$, $\mu^{(i)}(\cdot) \leftarrow 0$, $\epsilon \leftarrow 10^{-2}$,
2: **Exploration:** Set $\hat{W}^{(i)}(\cdot)$ and $\hat{L}^{(i)}(\cdot)$ according to simulation guidelines in §3.4.3.
3: **Exploitation:** Set $d(\cdot) = L^* - \hat{L}^{(i)}(\cdot)$
4: if $||d^{(i)}||_{[t_0,T]} < \epsilon$ then
5: $\mu_{SBRL}^*(\cdot) \leftarrow \mu^{(i)}(\cdot)$ stop
6: else
7: $\tau^{(i)}(t) = \min\{s | \hat{W}^{(i)}(s) < s - t\}, \ 0 \leq t \leq T$
8: $\mu^{(i+1)}(t) = \mu^{(i)}(t) + \frac{1}{\tau^{(i)}-t} \int_t^{\tau^{(i)}} \frac{d^{(i)}(a)}{W^{(i)}(s)} ds, \ 0 \leq t \leq T$
9: end if
10: $i = i + 1$ go to 2.

Algorithm Development

**Initialization:** $\mu^{(0)}(t)$

For simplicity of implementation, the initial candidate charge function is assumed to be a constant across the time horizon. We have found that a constant initial candidate does not significantly slow the convergence speed to a time-varying solution. The effect of the choice of initial charge value is explored in §3.4.4.

**Exploration:** $\mu^{(i)} \rightarrow E[L^{(i)}]$  

Given a candidate charge rate function, we can conduct discrete event simulation to simulate a BSS.
under the proposed charge control. While simulation is merely an estimation of the average charge
levels and prone to statistical errors, the error is reduced through increasing the number of independent
sample paths. Thus, the error can be made arbitrarily small through enough computational effort and we
assume a true value for average performance is obtained. See §2.4.3 for the simulation procedure used
to implement the examples in this work.

Exploitation: \( (\mu^{(i)}, \mathbb{E}[L^{i}]) \rightarrow \mu^{(i+1)} \)

Our objective is to achieve average HOL charge stabilization to some \( L^\star \) for a finite time-horizon \([0, T]\)
through an appropriate choice of charge rate function, \( \mu_{SBRL}^\star \). Our update procedure makes use of the
differences between \( \mathbb{E}[L^{i}] \) and \( L^\star \) to update the charge rate function. At time \( t \), we let the difference
between the realized HOL charge and the target charge level \( L^\star \) be \( d(t) \), where

\[ d(t) = L^\star - L^{i}(t) = L^\star - \int_{t-\mathbb{E}[W^{i}(t)]}^{t} \mu(s) ds. \] (3.78)

\( d(t) > 0 \) indicates the HOL battery has received and excess of charge while \( d(t) < 0 \) indicates the
battery has received a shortage of charge. The HOL battery at time \( t \) has been charging for an average of
\( \mathbb{E}[W^{i}(t)] = w_{t} \) time units. To achieve the desired HOL charge of \( L^\star \) at time \( t \) the amount of charge the
battery receives over its expected charging lifetime, \([t - w_{t}, t]\), must change by \( d_{t} \) units. We assume a
potential constant shift of the charge rate \( \Delta \mu_{t} = d_{t}/w_{t} \) across the entire charging interval. This ensures
the proper amount of charge has been distributed so that the resulting HOL of charge at time \( t \) is \( L^\star \). An
augmented charge rate function is created according to

\[
\tilde{\mu}(s, t) = \begin{cases} 
\mu(s) + \Delta \mu_{t}, & a_{t} \leq s \leq t \\
\mu(s), & \text{o.w.}
\end{cases} \] (3.79)

where \( \tilde{\mu}(t, s) \) is the charge rate needed at time \( s \) to achieve a HOL charge level of \( L^\star \) at time \( t \). We have
the augmented HOL charge \( \tilde{L}^{i} \) is equal to \( L^\star \)

\[ \tilde{L}^{i} = \int_{a_{t}}^{t} \mu(s) + \Delta \mu_{t} ds = L^\star. \] (3.80)

Figure 3.17 gives a graphical representation of the potential change in charge rate needed to achieve \( L^\star \)
at time \( t \). The gray shaded area indicates the charging interval of the HOL battery at time \( t \) and the blue
shaded area represents the increase in charge rate needed.

However, while this update achieves a pointwise stabilization at \( t \), in its current form it does not take
into account the changes in rate required to achieve stabilization across the entire interval. \( \tilde{\mu}(t, s) \) and
\( \mu(t + \delta, s) \) will have potentially different required rates over the common charging interval \([t + \delta - w_{t + \delta}, t]\). To achieve complete stabilization we must take into account all overlapping information. We
focus on the charge rate at time \( t \) to determine what information is pertinent to update its value. The
charge rate at time $t$ affects the charge of all batteries currently in the system. Therefore, the charge rate at time $t$ will in part determine the HOL charge level until the last batteries present in the system at time $t$ have left the system. We let $\tau(t)$ be the time at which the effect of charge level at time $t$ no longer determines the system charge state. This occurs when the HOL arrival falls after time $t$. We have

$$\tau(t) = \min\{s|w(s) < s - t\}. \quad (3.81)$$

Therefore, when choosing an appropriate new charge rate for time $t$, it is important to consider all system information on the interval $[t, \tau(t)]$. Figure 3.18 gives a graphical representation of the interval of interest and highlights the deviation from target level $L^*$. From the figure, we can see that the difference from $L^*$ is varying and, as a result, the augmented charge rate function, $\tilde{\mu}$, will have updated values for time $t$. We implement a simple time-average of all augmented charge rates across the interval. This will ensure system information at all time points affected by time $t$ will be considered.

$$\mu^{(i+1)}(t) = \mu^{(i)}(t) + \frac{1}{\tau(t) - t} \int_t^{\tau(t)} \tilde{\mu}(s,t)ds = \frac{1}{\tau(t) - t} \int_t^{\tau(t)} \frac{ds}{w_s} ds. \quad (3.82)$$

Termination: $\mu^{(i)} \rightarrow \mu^*_{SBRL}$

The goal of the SBRL algorithm is to achieve stabilization of the expected HCL to $L^*$. The error is captured by $d(t)$. A natural stopping criteria is when the maximum deviation from the target value across the time horizon is below some desired $\epsilon$ tolerance. The tolerance can be thought of as the percent difference acceptable in performance quality. Due to initial battery charges at the beginning of the time-horizon, it may not be feasible to achieve stabilization on the interval $[0, t_D]$. Therefore, the

Figure 3.17: Graphical representation of change needed in charge rate function, $\mu$, to achieve an expected HCL of $L^*$
maximum deviation is taken over the interval \([t_D, T]\) and the algorithm stops when it falls below \(\epsilon\). Although not stated in the algorithm, an alternative stopping criteria could be when successive iterations of the charge rate function fail to yield significant updates. That is when \(|\mu^{(i)} - \mu^{(i-1)}|_{[t_D,T]} < \epsilon\mu\) the algorithm is not making significant updates to the charge rate function and further iterations will not be prudent. This termination criteria does not guarantee any information of the resulting expected HCL but can be possibly useful in conjunction with the primary stopping criteria.

**Implementation Details**

Although the resulting control function \(\mu^*_{SBRL}(t)\) is desired to be a continuous-time function, discrete-event computer simulation necessitates the approximation of continuous time by dividing the time horizon into sufficiently small intervals of length \(\Delta\). Thus, the resulting charge rate is piecewise constant. Mainly,

\[
\mu^*_{SBRL}(t) = \mu^*_k, \quad k = \left\lfloor \frac{t}{\Delta} \right\rfloor. 
\]

In all implementations of this algorithm discussed in this chapter, we use \(\Delta = 0.01\). In addition, we capture the continuous stochastic performance functions of the system, such as HCL, TCL, and charge distribution, at finite intervals of length \(\Delta\). For any random process \(X(t)\), the data used to capture the random behavior is denoted as \(X^{(i)}_{n,k}\), where \(n\) is the index of the sample path, \(k\) is the time step index, and \((i)\) refers to the \(i^{th}\) iteration of the SRL algorithm. The Monte Carlo estimators over a time-horizon

---

**Figure 3.18:** Graphical representation of interval in which the charge rate at time \(t\) affects the HCL.
for a given iteration are calculated as

$$
\hat{X}_k^{(i)} = \frac{1}{N_S} \sum_{i=1}^{N_S} X_{n,k}^{(i)}, \quad 0 \leq k \leq \left\lfloor \frac{T}{\Delta} \right\rfloor.
\quad (3.84)
$$

where $N_S$ is the total number of independent sample paths generated for an estimator. We set $N_S$ to be 2000 samples for our experiments. Continuous charging of a battery is approximated as $L_k^{(i)} = L_{k-1}^{(i)} + \mu_{k-1}^{(i)} \Delta$.

### 3.4.4 Numerical Examples

In this section, we conduct numerical examples to test the effectiveness of the SBRL algorithm and provide a simultaneous comparison to the fluid algorithm when applicable.

**Base Case**

We return to the base case considered in the numerical experimentation of Algorithm 1 where customers bring fully depleted batteries to the system. We study a BSS with 50 available batteries. The sufficiently large scale of the system facilitates a valid fluid approximation so as to better compare with the SBRL method. We arrive customers to the system according to a NHPP with rate $\lambda(t) = 83.3 + 50 \sin(t)$. To clearly see the performance stabilization, a target HCL of 80% was chosen. $\mu_F^*$ was calculated according to Proposition 3. Algorithm 2 was implemented with $\epsilon = 0.01$, $N_S = 2000$, and $\mu_0 = 0$ to set $\mu_{SBRL}^*$. The results of the stabilization are presented in Figure 3.19 and we see excellent stabilization by both methods and $\mu_F^* \approx \mu_{SBRL}^*$. Achieving both the desired service quality and apparent agreement with the fluid method provides strong evidence for the validity of the SBRL algorithm. We show the convergence of both the HCL and $\mu_{SBRL}^*$ as well as the maximum and average error experienced in each iteration in Figure 3.20 and 3.21, respectively. With the parameters chosen, the SBRL algorithm terminated in 11 iterations. As we can see from comparing Figure 3.20 and 3.21, a small change in charge rate function can have a large impact on the expected HCL value. The first iterations of $\mu_{SBRL}^*$ do not visually appear to be changing from iteration to iteration. However, the difference in initial iterations of $L^\uparrow$ is clearly fluctuating in response.

**Small System Scale**

Because the SBRL algorithm is scale independent, it will continue to provide desired results even when the BSS has only a small inventory, thus complementing an area of weakness in Algorithm 1. To highlight both the difference in methods and the continued success of our SBRL algorithm, we modify the parameters of the previous example so the number of batteries in inventory is 2 and the arrival rate is $\lambda(t) = 3.33 + 2 \sin(t)$. Due to the smaller system size, the variance of the HCL is larger and, as a result,
a larger sample size is needed to achieve the same approximation precision as the previous example. However, in an effort to reduce computation requirements due to large number of sample runs, \( \epsilon \) was instead set to be 0.02 to allow for more stochastic volatility. Figure 3.22 shows that while both methods achieve apparent stabilization in all performance metrics, the SBRL algorithm is the only one effective at stabilizing the HCL to that target \( L^* = 0.8 \). The fluid method underestimates the charge required at all values. The SBRL algorithm was able to converge in 20 iterations.

**Residual Charge**

In this section, we include the important realistic feature of residual charge on incoming batteries. We modify the base case example to have customers enter the BSS with random residual charge. We present the simulation results under two choices of residual distribution. The first distribution is chosen to be a lognormal distribution with parameters chosen so that the results average and standard deviation for
Figure 3.20: Convergence of HCL under the SBRL Algorithm for the Base Case \((\lambda(t) = 83.3 + 50 \sin(t)\) and No Residual Charge)

Figure 3.21: Convergence of Charge Rate Function under the SBRL Algorithm for the Base Case \((\lambda(t) = 83.3 + 50 \sin(t)\) and No Residual Charge)
residual charge were 0.3 and 0.005, respectively. This distribution was chosen to approximate the empirical data presented in §3.2.1. Figure 3.25 gives the graphical comparison of empirical data and chosen distribution. While this distribution provides an approximate fit to actual data, the resulting distribution has small variance. To further test the robustness of our algorithm to a residual charge that is highly variable, we give the simulation results underneath a residual charge following a uniform distribution with bounds of 10% and 50%. We will only test the SBRL algorithm as it is the only algorithm proposed that will handle residual charge being present in the system. Using a tolerance of $\epsilon = 0.01$, the SBRL algorithm converges in only 9 iterations for the lognormal case and 11 iterations in the uniform case. Not only does the SBRL perform well at stabilizing the HCL, the results show that the method is relatively insensitive to variability of residual charge. However, as expected increasing the randomness of the incoming charge will increase the computation effort to achieve stable performance. We give the stabilization and convergence results in Figures 3.26 - 3.28 for the lognormal distribution and Figures

Figure 3.22: Comparison of SBRL and Fluid Algorithm Performance Stabilization of HCL, TCL, and Proportion 50% charged: Small Scale Case ($\lambda(t) = 8.3 + 5 \sin(t)$ and No Residual Charge)
Figure 3.23: Convergence of HCL under the SBRL Algorithm for the Small Scale Case ($\lambda(t) = 8.3 + 5 \sin(t)$ and No Residual Charge)

Figure 3.24: Convergence of Charge Rate Function under the SBRL Algorithm for the Small Scale Case ($\lambda(t) = 8.3 + 5 \sin(t)$ and No Residual Charge)
High QoS Target

Our model formulation allows for customers to take batteries from the BSS that have less than a full charge. However, a reasonable quality of service target is to stabilize expected HCL to approximately fully charged, thus high Qos target. We modify the small scale example to let $L^* = 0.95$. As stated in §3.4.2, at sufficiently small scales, Algorithm 1 will underestimate the charge required to reach fully charged. However, the SBRL algorithm is updated directly from the simulation output, thus taking into account the one-sided error from target value. All sample paths have a HCL battery level less than or equal to fully charged. The SBRL algorithm will update the charge until the negative error within an $\epsilon$ tolerance set. We use $\epsilon = 0.05$ for the simulation example presented in Figure 3.32. The epsilon is set to be higher to limit computational effort. Algorithm 2 terminated in 18 iterations. As expected, Algorithm 1 provided a much smaller charge rate than required to truly stabilize the mean HCL at the desired high QoS target. The SBRL algorithm terminates when the maximum difference from target is within $\epsilon$ distance.

The charge rate function found to achieve fully charged HOL batteries is not unique. In fact, there exist an infinite amount of variations that will provide sufficient energy to the system. We are interested
in the minimum charge rate function possible that will still obtain desired charges. We submit that the SBRL output provides a reasonable approximation of this minimum charge. In theory, the averaging update process of the SBRL has the potential to over correct and yield a HCL higher than the target level. Because $L^* = 1$ represents a boundary, the overcharging would not be captured and the algorithm would not be able to correct itself. Thus, SBRL potentially does not yield the minimum. However, the update process is also based on the expected HCL level. As discussed, the expectation will only experience negative deviation from 1 if that is the target value. Thus the update procedure will rarely experience a value of 1 for the expected HCL refinement. Because the overcharging will not occur in algorithm implementation, SBRL provides a reasonable charge rate function that can achieve approximately fully charged batteries at a minimum charge rate.

The charge rate function found to stabilize the expected HCL gives the lower bound on charge that needs to be added to the system in order to fully charge batteries to meet demand. Using this lowerbound,
Figure 3.27: Convergence of HCL under the SBRL Algorithm for the Lognormal Residual Case

Figure 3.28: Convergence of Charge Rate Function under the SBRL Algorithm for the Lognormal Residual Case
we can add an engineering refinement for potential alternate objective satisfaction. For instance, high rates of charge are undesirable as they can damage the battery and reduce its life expectancy [92]. In order to take this into account, if the peak charge rates are undesirably high, a system manager can choose to increase the charge rate in the valley charge rate time intervals to charge batteries up to desired levels before peak demand at a slower rate. This will be effective as long as the area under the modified charge rate control is greater than or equal to the area under the curve of the charge rate given.

**SBRL Robustness to Initial Candidate Value**

In this experiment, we test the robustness of the SBRL algorithm to the initial candidate function value starting above the desired charge rate function. In all previous examples, we have set the initial candidate value to be 0, which is guaranteed below the desired charge rate at all time points. For this example, we will use the previous example parameters except for setting the initial candidate charge value $\mu^0 = 1.5$,
Figure 3.30: Convergence of HCL under the SBRL Algorithm for the Uniform Residual Case

Figure 3.31: Convergence of Charge Rate Function under the SBRL Algorithm for the Uniform Residual Case
Figure 3.32: Simulation Results for SBRL Algorithm Performance Stabilization of HCL, TCL, and Proportion Charged at Least 50% with $L^* = 0.95$

which is greater than the desired charge rate throughout the time horizon. The convergence takes longer than the base case at 14 iterations. Interestingly, the shape of successive iterations is more distinct than the base case as seen in Figure 3.35. The base case achieved the relative shape of the final charge control function after approximately 3 iterations while this case has a gradual descent to the final form. The algorithm appears to better perform when operating under a candidate function that has a deficit from the desired charge. We believe this is due to the relative difference in wait times at the peak charge times. The greater error in the $d(t)$ function is now coupled with a larger interval of influence, thus leading to slower convergence.
3.5 Concluding Remarks

In this chapter, we created a stochastic queueing model representation of a BSS facility that focuses on capturing dynamic charge levels of inventory batteries as a function of time-varying customer demand. Motivated to capture both continuous charge levels over a continuous time horizon, expressions were developed for key performance indicator functions, such as highest available battery charge, total charge inventory in the system, and quantity of batteries charged to a quality level. The model was extended to include such realistic features as non-linear charging and random residual battery charges. We proposed and defined representative fluid models to capture the BSS dynamics and provide continuous, deterministic approximating functions for average system behavior. Numerical results show that our proposed functions effectively capture average behavior in representative examples. Building upon these results, we developed a charge rate control policy in an effort to stabilize the quality of service (battery charge received) experience by an incoming customer. To solve for the required rate control policy, both a fluid
Figure 3.34: Convergence of HCL under the SBRL Algorithm with Initial Candidate Value $\mu^0 = 1.5$

Figure 3.35: Convergence of Charge Rate Function under the SBRL Algorithm with Initial Candidate Value $\mu^0 = 1.5$
model approach and SBRL algorithm were created. Through numerical experimentation, both methods showed excellent ability to find a control function that achieved stabilization. These performance analysis and control techniques provide promising basis for future study into optimal BSS control.
Chapter 4

Conclusions and Future Work

4.1 Dissertation Summary and Contributions

Service systems experiencing time-varying demand require careful consideration to ensure that the system is properly analyzed and managed. Slight changes in operation of a system can lead to periods of vastly different customer experiences and resulting quality of service. Thus, any potential control policy must exhibit similar time-varying behavior to adequately achieve desired results. In customer service systems, a customer’s quality of service is a key metric in evaluating the success of a business. This dissertation focuses on building tools and methods to develop control policy recommendations that will ensure a stabilized quality of service is experienced by customers even in the face of time-varying demand. This work contributes to the field of Operations Research in both stochastic service management and queueing theory.

In Chapter 2, we studied a time-varying multiclass service system with differentiated service requirements. Our goal was to stabilize the TPoD for each class to class-dependent targets. We developed joint staffing and scheduling policies that were proven using heavy-traffic diffusion limits to asymptotically achieve performance stabilization to the desired targets. These control policies are both time dependent and state dependent (based on real-time customer delays). By incorporating the real-time delay into our control policy decision, we were able to control not only the average system behavior but stochastic fluctuation of the system as well. Extensive simulation experiments were conducted and showed the effectiveness of our proposed policies in a wide range of model settings.

In Chapter 3, we studied a BSS with time-varying arrivals and residual battery charges. We established a novel representation of inventory charge levels through the use of a queueing representation and resulting wait times. Fluid model approximations were developed and shown through numerical examples to effectively capture inventory charge levels. A centralized charge control policy was proposed to stabilize the resulting battery charge presented to an incoming customer. The charge control was found through both a fluid approximation algorithm and a simulation-based reinforcement learning algorithm.
Numerical examples show both algorithms are effective at stabilizing inventory charge levels.

We summarize the contributions of this dissertation below.

1. Proposal of staffing and scheduling policies for a multiclass model that incorporates both time-varying arrivals and dynamic, real-time system information.

2. Specification of analytical control functions to achieve stabilized and differentiated tail-probability of delays in a multiclass model through the establishment of many-server heavy-traffic limit theorems.

3. Creation of novel stochastic queueing systems and fluid models to approximate inventory charge levels of a battery swapping station.


4.2 Future Work

Both chapters of work presented in this dissertation give rise to intriguing possible extensions and opportunities for further refinement of the analysis and control of time-varying service systems.

Alternate V Model Control

Chapter 2 gives a staffing and scheduling policy based on HOL waiting times. In practical implementation, this requires a method for storing and updating waiting times of all customers currently in the system. This creates a high storage and information retrieval cost on the system. The storage requirements scale with the total number of customers in the system. To reduce the dimension of information needed to make a scheduling and routing decision, we propose the utilization of summary count information for each customer class (i.e. queue lengths and number of allocated servers). This will reduce the dimensionality of required storage to $K$, where $K$ is the number of classes in the system. By attempting to control the total server allocation to each target class, we are essentially splitting the multiclass system into $K$ parallel queues operating at class-specific service capacities. The queue-length based routing policy can potentially be achieved by translating the required wait time into necessary queue lengths. The potential difference in total service capacities required by the three proposed staffing and scheduling methods (delay, service, and queue based) motivates their study in an effort to determine which policy yields the least number of servers to achieve system stabilization.
SBRL Convergence

Chapter 3 proposed a novel stochastic system and control for a BSS. The directions for this work are twofold. First, the fluid models and SBRL algorithm developed can be further analyzed to develop sound mathematical results. While shown to be extremely effective in simulation examples, the SBRL algorithm can be further strengthened by providing a proof of algorithm convergence. We can then be assured that our algorithm will give meaningful results in all model cases. In addition, proving convergence of this reinforcement learning method would provide the basis for similar techniques being applied to a wide range of other performance stabilization problems.

BSS Cost Optimization

Our preliminary work on BSS analysis and control has been focused on performance stabilization. Although not explicitly stated previously, we can formulate the problem studied in terms of a stochastic optimization problem with the following objective

\[
\min_{\mu > 0} \left| E[L^\uparrow] - L^* \right|_T,
\]  (4.1)

where \( \left| \cdot \right|_T \) indicates the largest absolute magnitude over an interval \([0, T]\). This objective function can be coupled nicely with additional optimization objectives. To fully capture all operational goals of a BSS, we wish to merge our quality of service stabilization with a cost optimization objective. By incorporating both a need for high quality of service and cost-efficient operating procedures, we can provide a holistic approach to better operation of a BSS. To supplement our service quality performance objective, we give a possible formulation of a mathematical model focused on operating costs. We give the optimization framework using the stochastic model and charging methods presented in Chapter 3. While a majority of the literature considers a cost minimization objective, see [92] and [89] for examples, we propose a profit maximization structure due to our introduction of a revenue stream. We assume that incoming customers will pay for the differential in energy between the batteries being swapped (i.e. the HCL and the random residual charge).

For our formulation, we let \( C_p \) be the price per energy unit differential the customer pays to undergo a battery swapping. Similar to [89] we allow the cost of electricity to be time-varying to reflect fluctuating energy markets. We let \( p(t) \) be the cost per unit energy at time \( t \). We can describe the total energy consumption of the BSS over a time horizon of length \( T \) as

\[
P_T(\mu) = \sum_{i=1}^{N(T)} C_p \cdot (L^\uparrow(\eta_i) - \gamma_i)^+ - \int_0^T p(t) \mu(t) B(t, 1) dt,
\]  (4.2)

\[
= C_p \int_0^T \int_0^1 (L^\uparrow(t) - F^{-1}(x))^+ dU(N(t), x) - \int_0^T p(t) \mu(t) B(t, 1) dt,
\]  (4.3)

130
where $B(t, 1)$ is the number of batteries less than full and actually receiving charge. We wish to find the charge rate function $\mu$ that will maximize the expected total profit over the time horizon.

$$\max_{\mu > 0} \mathbb{E}[P_T(\mu)].$$ (4.4)

This is essentially an optimal control problem. The control variable is the continuous charge rate function, $\mu$. We have two state variables of importance to our optimization, the expected HCL, $\mathbb{E}[L^+(t)]$, and the expected number of fully charged batteries, $\mathbb{E}[B(t, 1)]$. However, standard tools of deriving optimal control (such as the Hamilton-Jacobi-Bellman (HJB) and Pontryagin’s Principle) require a explicitly defined function of the change in state variable with respect to the change in control. Our problem does not have such a clear relationship between system state and control.

We believe a promising direction forward is the use of simulation optimization to find a desired control function, mainly a stochastic gradient descent (SGD) method. Similar to the SBRL concept introduced in Chapter 3, we can use SGD by iteratively conducting computer simulations to inform an smart update procedure. The SGD relies on the gradient of the objective function with respect to the control functions. Mainly,

$$\mu^{(i+1)} = \mu^{(i)} + \eta^{(i)} \hat{H}_i(\mu^{(i)}),$$ (4.5)

where $\eta^{(i)}$ is an iteration-dependent step size or learning rate and $\hat{H}_i$ is an estimation of the gradient of the objective function under $\mu^{(i)}$ control. We wish to first develop an SGD method for the objective function in (4.1) and validate it by comparing to the charge rate function found by our current SBRL method. Then, utilizing the established techniques, we wish to fully incorporate the cost structure proposed in (4.2) to determine more practical BSS operating procedures.
REFERENCES


APPENDICES
Appendix A

Appendix for Chapter 2

A.1 Multiclass Service System without Abandonment

Because customer abandonment is negligible in some service systems (e.g., health care systems), we provide a study of the multiclass V model presented in Chapter 2 under the assumption of infinitely patient customers (i.e. no abandonment). The decision to model abandonment is an important consideration to determine true system behavior and by providing mathematical support and discussion of the no-abandonment case, we can provide a more complete multiclass modeling framework. As we will see, the impact of removing abandonment from the model structure has a fundamental impact on the results, both mathematically and experimentally. In the following sections, we provide corollaries of Chapter 2 results that yield control policies. Numerical experimentation is conducted and the resulting implications on method practicality are discussed. Without customer abandonment, the first order staffing process becomes

\[ \bar{m}_i(t) = \int_0^t e^{-\mu_i(t-u)} \lambda_i(u - w_i) du. \]  

(A.1)

Similarly, define \( \bar{q}_i(t) = \int_{t-w_i}^t \lambda_i(u) du \). The MSHT scalings are similar to those in 2.3.1, except that we now use \( \bar{m}_i \) and \( \bar{q}_i \) as the centering components in place of \( m_i \) and \( q_i \).

A.1.1 Many-Server Heavy Traffic Functional Central Limit Theorem

Corollary A.1.1 (MSHT FCLT limits without abandonment) Suppose that the conditions of Theorem 2.3.1 are satisfied and customers are infinitely patient. Then there is a joint convergence for the CLT-scaled processes.

\[
\left( \hat{H}_n, \tilde{B}_1^n, \ldots, \tilde{B}_K^n, \hat{H}_1^n, \ldots, \hat{H}_K^n, \tilde{V}_1^n, \ldots, \tilde{V}_K^n, \tilde{X}_1^n, \ldots, \tilde{X}_K^n, \tilde{Q}_1^n, \ldots, \tilde{Q}_K^n \right)
\Rightarrow \left( \hat{H}, \tilde{B}_1, \ldots, \tilde{B}_K, \hat{H}_1, \ldots, \hat{H}_K, \tilde{V}_1, \ldots, \tilde{V}_K, \tilde{X}_1, \ldots, \tilde{X}_K, \tilde{Q}_1, \ldots, \tilde{Q}_K \right)
\text{ in } D^{5K+1}, \quad n \to \infty,
\]
as $n \to \infty$, where the FCLT limits on the right-hand side are well-defined stochastic processes.

(i) The limiting processes $(\hat{H}, \hat{B}_1, \ldots, \hat{B}_K)$ jointly satisfy the following set of $K$ Ornstein-Uhlenbeck (OU) type stochastic integral equations

$$\hat{B}_i(t) + \eta_i^0(t) \hat{H}(t) = - \int_0^t \mu_i \hat{B}_i(u) du + \eta_i^0(t) \kappa_i(t) + G_i(t)$$

for $i=1, \ldots, K$, and $\sum_{i=1}^K \hat{B}_i(t) = c(t)$, where

$$\eta_i^0(t) \equiv w_i \lambda_i (t - w_i), \quad G_i(t) \equiv \hat{E}_{i,1}(t) - \hat{D}_i(t),$$

$$\hat{E}_{i,1}(t) \equiv \int_{-w_i}^{t-w_i} \sqrt{\lambda_i(u)} dW_{\lambda_i}(u), \quad \hat{D}_i(t) \equiv \int_0^t \sqrt{\mu_i \mu(u)} dW_{\mu_i}(u),$$

and $W_{\lambda_i}, W_{\mu_i}$ are independent standard Brownian motions.

(ii) The FCLT limits for all HWT and PWT processes are deterministic functionals of a one-dimensional process $\hat{H}$, namely,

$$\hat{H}_i(t) \equiv w_i (\hat{H}(t) - \kappa_i(t)), \quad \text{and} \quad \hat{V}_i(t) = w_i (\hat{H}(t + w_i) - \kappa_i(t + w_i)).$$

(iii) The FCLT limit for each queue-length process is the sum of three terms, namely, $\hat{Q}_i(t) \equiv \hat{Q}_{i,1}(t) + \hat{Q}_{i,2}(t)$, where

$$\hat{Q}_{i,1}(t) \equiv \int_{t-w_i}^{t} \sqrt{\lambda_i(u)} dW_{\lambda_i}(u), \quad \hat{Q}_{i,2}(t) \equiv \lambda_i(t-w_i) \hat{H}_i(t).$$

(iv) Finally, the FCLT limits for number in system is given by $\hat{X}_i(t) = \hat{B}_i(t) + \hat{Q}_i(t)$.

Proof of Corollary A.1.1. The proof is very similar to that of Theorem 2.3.1. Step 1 remains unchanged as the relations (2.47) and (2.49) are unaffected by abandonment. For Step 2, instead of using the decomposition in (2.51), we break the enter-service process $E^n_i(t)$ into two terms, namely,

$$E^n_i(t) \equiv \sqrt{n} \int_{-w_i}^{t-H^n_i(t)} d\hat{A}_i^n(u) + n \int_{-w_i}^{t-H^n_i(t)} \lambda_i(u) du.$$
Accordingly, we obtain (paralleling (2.57) and (2.61))

\[
\hat{E}_i^n(t) = \int_{-w_i}^{t-H_i^n(t)} d\hat{A}_i^n(u) + \sqrt{n} \int_{-w_i}^{t-H_i^n(t)} \lambda_i(u) du
\]

where \(\zeta_i^n(t)\) satisfies (2.62). Step 3 follows closely that of Theorem 2.3.1. The convergence of \(\hat{E}_{i,1}^n \Rightarrow \hat{E}_{i,1}\) holds as it is the scalar multiple of the convergent process in (2.56). The departure process convergence is unaffected by abandonment. The arguments from (2.63) up to (2.72) are still valid. This allows us to conclude

\[
\hat{B}_i^n(t) + w_i \lambda_i(t - w_i) \hat{H}(t) = -\mu_i \int_0^t \hat{B}_i^n(u) du + \int_{-w_i}^{t-w_i} d\hat{A}_i^n(u) + O(n^{-1/2})
\]

from which the joint convergence of \((\hat{H}^n, \hat{B}_1^n, \ldots, \hat{B}_{K-1}^n)\) follows by application of the continuous mapping theorem and joint convergence of \(\hat{E}_{i,1}^n\) and \(\hat{D}_{i}^n\).

### A.1.2 Distribution of Frontier Process

Similar to the case with customer abandonment, we can derive the distribution of \(\hat{H}\) and use it to develop the desired staffing and scheduling policies.

**Corollary A.1.2 (Distribution of the frontier process \(\hat{H}\) without abandonment)** The process \(\hat{H}\) uniquely solves the following stochastic Volterra equation (SVE)

\[
\hat{H}(t) = \int_0^t L(t, s) \hat{H}(s) ds + \int_0^t J(t, s) dW(s) + K(t), \quad (A.3)
\]

where

\[
L(t, s) = \sum_{i=1}^K \eta_i^a(t) e^{\mu_i(s-t)} \mu_i, \quad J(t, s) = \frac{\sqrt{\sum_{i=1}^K e^{2\mu_i(s-t)} (\lambda_i(s - w_i) + \mu_i m_i(s))}}{\eta_i^a(t)},
\]

\[
K(t) = \sum_{i=1}^K \left( \eta_i^a(t) \eta_i^c(t) - \mu_i \int_0^t \eta_i^a(s) e^{\mu_i(s-t)} \eta_i^c(s) ds \right) - c(t) \quad (A.4)
\]

\(W\) is a standard Brownian motion. In addition, \(\hat{H}\) is a Gaussian process with

(i) mean \(M_{\hat{H}}(t) = E[\hat{H}(t)], 0 \leq t \leq T,\) uniquely solving the fixed-point equation (FPE)

\[
M_{\hat{H}} = \Gamma(M_{\hat{H}}), \quad \text{where} \quad \Gamma(M_{\hat{H}})(t) = \int_0^t L(t, s) M_{\hat{H}}(s) ds + K(t), \quad (A.5)
\]

146
(ii) covariance $C_{\tilde{H}}(t, s) \equiv \text{Cov}(\tilde{H}(t), \tilde{H}(s)), 0 \leq s, t \leq T$, uniquely solving the FPE

$$C_{\tilde{H}} = \Theta(C_{\tilde{H}}).$$

where the operator $\Theta$ is defined as

$$\Theta(C_{\tilde{H}})(t, s) \equiv -\int_0^t \int_0^s L(t, u) L(s, v) C_{\tilde{H}}(u, v) dv du + \int_0^t L(t, u) C_{\tilde{H}}(u, s) du + \int_0^s L(s, v) C_{\tilde{H}}(t, v) dv + \int_{s \wedge \mu}^s J(t, u) J(s, u) du. \quad (A.6)$$

**Proof of Corollary A.1.2** The multi-dimensional SDE (A.2) is equivalent to

$$\frac{d}{dt} \left( e^{\mu t} \tilde{B}_i(t) \right) = e^{\mu t} \left( -\eta_i^a \tilde{H}(t) - \eta_i^a \kappa_i(t) + G_i(t) \right), \quad (A.7)$$

Following similar steps as Corollary 1, we can sum over all classes and utilize the relationship

$$\sum_{i=1}^K \tilde{B}_i(t) = c(t)$$

to achieve a convenient expression of just the frontier process. Differentiating the resulting equation gives the SVE in A.3. To show uniqueness, we define a similar operator, $\psi$, as in (2.86) but using $L$ as defined in Corollary A.1.2. We again have $\|\psi(x_1) - \psi(x_2)\|_T \leq L^\uparrow T \|x_1 - x_2\|_T$, but where the bounding constant is expressed as

$$L^\uparrow = \frac{\sum_{i=1}^K w_i \lambda_i^\uparrow \mu_i}{\sum_{i=1}^K w_i \lambda_i^\uparrow F_i^c(w_i)} < \infty. \quad (A.8)$$

The bound is guaranteed by the strict positivity assumptions on $w_i$, $\lambda_i$ and $F_i^c$ for all $1 \leq i \leq K$. A partitioning argument along with Banach fixed point theorem implies that the FPE has a unique solution over $[0, T]$. This bound also guarantees the contraction property of $\text{Cov}(\tilde{H}(t), \tilde{H}(s))$. The proof of Gaussian result for $\tilde{H}(t)$ follows exactly as in Corollary A.1.2.

**A.1.3 Asymptotically Unique Control Functions**

**Corollary A.1.3 (Asymptotically unique control functions without abandonment)** The condition (2.32) is satisfied in a system without abandonment if and only if

$$c(t) = \sum_{i=1}^K \left\{ \eta_i^a(t) \kappa_i(t) - \int_0^t \eta_i^a(s) \mu_i e^{\mu_i(s-t)} \kappa_i(s) ds \right\}, \quad (A.9)$$

$$\kappa_i(t) = z_{1-\alpha_i} \sigma_{\tilde{H}}(t) + \Delta(t), \quad 1 \leq i \leq K. \quad (A.10)$$

where $\Delta$ is a finite function that can be common to all prioritization regulators. The proof follows the
exact structure of the main paper. As such, we omit restating the proof here.

A.1.4 Differentiation of the mean PWT

Corollary A.1.4 (Asymptotic service differentiation and performance stabilization without ab.)

Under TV-SRS (2.14) and TV-DPS (2.15) with $c_i(\cdot)$ and $\kappa_i(\cdot)$ specified in Corollary A.1.3 we have the following asymptotic stability results:

(i) Mean PWT and mean HWT are both asymptotically stabilized for all classes:

$$
\mathbb{E}[V^n_i(t)] \to w_i \quad \text{and} \quad \mathbb{E}[H^n_i(t)] \to w_i \quad \text{as} \quad n \to \infty, \quad 1 \leq i \leq K, \quad 0 < t \leq T.
$$

(A.11)

(ii) TPoDs for PWT and HWT are both asymptotically stabilized for all classes:

$$
\mathbb{P}(V^n_i(t) > w_i) \to \alpha_i \quad \text{and} \quad \mathbb{P}(H^n_i(t) > w_i) \to \alpha_i \quad \text{as} \quad n \to \infty, \quad 1 \leq i \leq K, \quad 0 < t \leq T.
$$

(A.12)

Proof of Corollary A.1.4

The FCLT limits in Theorem A.1.1 implies the FWLLN, that is, we have

$$(H^n_i, V^n_i) \Rightarrow (w_i, w_i) \quad \text{in} \quad \mathcal{D}^2, \quad \text{for} \quad 1 \leq i \leq K, \quad \text{as} \quad n \to \infty,$$

where $\epsilon(t) = 1$. To prove part (i) of Theorem A.1.4, it is sufficient to show that $\{V^n_i, n \geq 1\}$ and $\{H^n_i, n \geq 1\}$ are uniformly integrable (u.i.). We first prove that the queue length $Q^n_i$ is u.i. To do so, note that the queue length is necessarily bounded by the number of arrivals to the system $A^n_i$. We have $Q^n_i(t) \leq A^n_i(t)$. Because $A^n_i(t)$ is a Poisson r.v., the u.i. of $A^n_i(t)$ is straightforward. Specifically, we have

$$
\sup_n \mathbb{E} \left[ (A^n_i(t))^2 \right] = \sup_n \left[ \int_0^t \frac{\lambda^n_i(t-x)}{n} dx + \left( \int_0^t \frac{\lambda^n_i(t-x)}{n} dx \right)^2 \right] < \infty, \quad (A.13)
$$

Next, we write the PWT

$$
V^n_i(t) = \sum_{j=0}^{Q^n_i(t)} U_j,
$$

where $U_j$ is the time between the $j^{th}$ and $(j+1)^{th}$ departure times of existing waiting customers at queue $i$. $U_j$ is a summation of the random number of customers exponential service times that enter
before the $j^{th}$ customer. In Theorem 2.3.1 we argued that the number of customers to enter before the HOL class $i$ customer was bounded by a geometric distribution. Thus, similar to Theorem 2.3.2, we can bound the second moment of potential waiting time as

$$
\mathbb{E} \left[ V^n_i(t)^2 \right] = \mathbb{E} \left[ \sum_{j=0}^{Q^n_i(t)} (U_j)^2 + \sum_{j \neq k} U_j U_k \right] 
\leq (\mathbb{E}[Q^n_i(t)] + 1) \frac{\ell^{(1)}_i}{(nm^2 \tilde{\mu})^2} + \mathbb{E}[Q^n_i(t)^2 - Q^n_i(t)] \left( \frac{\ell^{(1)}_i + 1}{(nm^2 \tilde{\mu})^2} \right)
$$

where $\tilde{\mu} \equiv \min_{1 \leq i \leq K} \mu_i$.

Using the bound in (A.13), we have $\sup_n \mathbb{E} \left[ V^n_i(t)^2 \right] < \infty$, which implies u.i. of $V^n_i$. The u.i. of $H^n_i$ is straightforward because $0 \leq H^n_i \leq T$.

The TPoD for class-$i$ customers

$$
P(V^n_i(t) > w_i) = P(\sqrt{n} \left( V^n_i(t) - w_i \right) > 0) = P(\tilde{V}^n_i(t) > 0)
\rightarrow P(\tilde{V}_i(t) > 0) = P \left( \tilde{w}_i \left( \tilde{H}(t + w_i) - \kappa_i(t + w_i) \right) > 0 \right)
= P \left( \tilde{H}(t + w_i) > \kappa_i(t + w_i) \right) = P \left( Z > \frac{\kappa_i(t + w_i)}{\sigma_H(t + w_i)} \right) = P(\mathcal{Z} > \alpha_i) = \alpha_i,
$$

where the third equality holds by (A.2).

**Remark A.1.1** The results in Corollaries A.1.1-A.1.4 closely mirror the corresponding theorems and propositions in the main body of work. However, the equations presented no longer have a dependency on a Brownian motion term associated with the abandonment process ($\mathcal{W}_b$). The diffusion results now only depend on the randomness associated with the arrival and service processes. However, as we will see in following sections, the loss of abandonment stochasticity does not imply more stable results in the stochastic model. Customers abandoning the system has a load balancing effect that ensures that expected queue lengths and wait times do not grow excessively large. Without abandonment, customers are forced to wait in the system until completion. Thus, any deviation from the desired performance levels will affect all future customer experiences. As a result, we see in the following corollaries and numerical experimentation that the variance of the frontier process grows larger as more time evolves.

We also remark that for a set of model parameters that includes a mixture of classes that abandon and that do not, similar corollaries can be derived by carefully distinguishing the type of customer. For a model with $K$ classes, we let $m^*$ be the number of classes that allow for customer abandonment. Without loss of generality, we can number the abandonment classes from 1 to $m^*$. For classes $i = 1, \ldots, m^*$, we include the Brownian abandonment term in the derivation. For implementation, we propose utilizing the formulas in the main body of work while setting $f_i(x) = 0$ and $F_i^c(x) = 1$ for classes $i = m^*, \ldots, K$. 

---

149
A.1.5 Class Independent Service Rate

Corollary A.1.5 (Frontier process $\widehat{H}$ when service rates are class independent without ab.)
Suppose $\mu_i = \mu$, $1 \leq i \leq K$, then

(i) the limiting frontier process $\widehat{H}$ satisfies the one-dimensional OU type SDE

$$\eta^a(t)\widehat{H}(t) = -\int_0^t \eta^a(u)\widehat{H}(u) du + S(t) + G(t), \quad (A.14)$$

where $G(t) \equiv \sum_{i=1}^K G_i(t)$, for $G_i(t)$ being the Brownian-driven terms given in Corollary A.1.1, and

$$S(t) \equiv \sum_{i=1}^K \eta^0_i(t) \kappa_i(t) - c(t) - \mu \int_0^t c(u) du.$$

(ii) The SDE (A.14) has a unique solution

$$\widehat{H}(t) = \frac{1}{R(t)} \left( \int_0^t e^{\int_u^t \frac{L(v)}{R(v)} dv} \mathcal{J}(u) dW(u) + \int_0^t e^{\int_u^t \frac{L(v)}{R(v)} dv} R(u) dK(u) \right. \quad (A.15)$$

$$+ \left. \int_0^t e^{\int_u^t \frac{L(v)}{R(v)} dv} K(u) dR(u) \right), \quad (A.16)$$

where $W$ is a standard Brownian motion,

$$R(t) = e^{\mu t} \eta^a(t), \quad L(t) = \mu e^{\mu t} \eta^a(t), \quad \mathcal{J}(t) = e^{\mu t} \sqrt{\sum_{i=1}^K \frac{\lambda_i(t - w_i)}{\mu} + \mu m_i(t)}.$$

(iii) The variance of $\widehat{H}(t)$ is

$$\sigma^2_{\widehat{H}}(t) \equiv \text{Var} \left( \widehat{H}(t) \right) = \frac{2\Lambda_w(t) - m(t)}{\left( \sum_{i=1}^K \eta^0_i(t) \right)^2} \quad (A.17)$$

where

$$\Lambda_w(t) = \sum_{i=1}^K \Lambda_i(t - w_i), \quad m(t) = \sum_{i=1}^K m_i(t).$$

Remark A.1.2 Unlike the variance expression in Chapter 2, the result in Corollary A.1.5 indicates that the frontier process grow larger the longer the system is in operation. This makes sense intuitively as the presence of abandonment provides a self-correcting effect on queue lengths and system wait times. Without this stabilizing effect, the system will experience the effects of all customers who have visited the system.
Proof of Corollary A.1.5. The form of the unique solution to the frontier process is the same as given in 2.3.2. We further simplify the resulting variance using the simplified formulas derived for no-abandonment.

\[
R^2(t) \text{Var} \left( \hat{H}(t) \right) = \int_0^t e^{2 \int_u^t \frac{L(v)}{\Pi(v)} dv} \tilde{j}^2(u) du, \quad (A.18)
\]

\[
= \sum_{i=1}^K e^{2\mu t} \int_0^t \lambda_i(u - w_i) du + \mu e^{2\mu t} \int_0^t e^{-\mu u} \int_0^u e^{\mu s} \lambda_i(s - w_i) ds du, \quad (A.19)
\]

\[
= \sum_{i=1}^K e^{2\mu t} \Lambda_i(t - w_i) + \mu e^{2\mu t} \int_0^t e^{\mu s} \lambda_i(s - w_i) \frac{e^{-\mu t} - e^{-\mu s}}{-\mu} ds \quad (A.20)
\]

\[
= \sum_{i=1}^K 2e^{2\mu t} \Lambda_i(t - w_i) - e^{\mu t} \int_0^t e^{\mu s} \lambda_i(s - w_i) ds \quad (A.21)
\]

\[
\text{Var} \left( \hat{H}(t) \right) = \frac{\sum_{i=1}^K 2\Lambda_i(t - w_i) - m_i(t)}{(\sum_{i=1}^K \eta_i(t))^2} \quad (A.22)
\]

where the third equality holds from interchanging integrals and the fifth holds from the definition of \(m_i(t)\) defined under the assumption of no abandonment. □

A.1.6 Constant Arrival

Corollary A.1.6 (Constant arrival rates without abandonment) When \(\lambda_i(t) = \lambda_i\), we have

\[
m_i(t) \sim m_i \equiv \frac{\lambda_i}{\mu}, \quad e(t) \to 0 \quad (A.23)
\]

\[
\kappa_i(t) \sim \frac{z_1 - \alpha_i}{\sum_{j=1}^K \eta_j} \sqrt{2\lambda \Sigma t}. \quad (A.24)
\]

where \(\eta_i^a = w_i \lambda_i\), \(\lambda \Sigma = \sum_{i=1}^K \lambda_i\), and we say \(f(t) \sim g(t)\) if \(f(t)/g(t) \to 1\) as \(t \to \infty\).

Remark A.1.3 Because the variance of the frontier process grows with time, there does not exist a steady state for the prioritization regulator. However, the second order staffing term goes to 0 as time grows to be large. This indicates that as the system evolves, a safety stock of servers is no longer needed to stabilize TPoD. We have the required staffing to be equal to the first order staffing term, (i.e. \(s(t) = m(t)\)). For \(t\) large and finite \(n\), the prioritization regulators will overcome the real-time waiting information in the scheduling decision. Wait times will have to constantly increase as time evolves in order for the corresponding customer to be chosen to enter service. Furthermore, there exists a class \(i^\dagger\) that has a prioritization regulator \(\kappa_{i^\dagger}(t)\) that will dominate increasingly the other regulators. This ensures that as time grows, the wait times of the non-dominating classes will need to grow in order to enter service. This approaches an absolute prioritization model where if there is a customer
of the dominating class, they will be scheduled into service next, regardless of wait time. While our ultimate goal of stabilizing TPoD is shown to be effective in the following numerical section for a finite time-horizon and large system scale, the mathematical framework is built off of an assumption of an overloaded system (i.e. positive wait times and queue lengths). With the absolute prioritization and a small system scale, this assumption will be violated as dominating class customers will experience extremely small wait times and little queueing. However, we remark that as the scale of the system increases, the longer the system can operate before experiencing poor results. This is because of the order in which we performed our limits. Each stochastic model process can be indexed by both a system scale \( n \) and time \( t \), and thus two directions to follow to determine the steady state limiting model. We first determine the limiting model by letting \( n \) go to infinity and then attempt to determine the steady state control of the limiting system as \( t \) goes to infinity. Therefore, we can increase the performance of our method at large \( t \), by increasing the system scale as shown in Corollary A.1.4. The other ordering of taking limits (i.e. determining steady state of the stochastic system and then letting the scale grow) remains open work.

### A.1.7 Numerical Results

In this section, we provide numerical testing of the no abandonment case under several conditions. The lack of abandonment stabilization slightly reduces the effectiveness of our approach.

**Base Case without Abandonment**

We modify the base case example in the main work to make customers infinitely patient. We still achieve desired quality of service levels as seen in Figure A.1. Class 1 TPoD appears to fluctuate periodically around the target level. We explore further the behavior of the waiting queues under our proposed control. While TPoD is effectively stabilized, Figure A.2 shows that for a finite \( n \), the average customer experience diverges from the target levels. Class 1 appears to achieve higher prioritization of the two classes as time evolves. Both the potential delay and queue length for Class 1 drifts to zero. In the bottom panel, we see that the probability of an empty Class 1 queue increases as time increases. In contrast, the wait times and queue lengths for Class 2 grow over time. We submit that the slight periodic fluctuation observed in the TPoD is due to the increasingly extreme behavior.

The manner in which we control the TPoD perturbs the expected waiting at finite scale so that the TPoD can be achieved. We know from a Little’s Law relationship that the wait times are directly related to the queue lengths. However, without the self-correcting effect of abandonment, the required modification to achieve second-order control will have a compounding effect on the system. The results of our diffusion based scheduling and staffing rely on the assumption that the multiclass model will experience positive queue lengths and wait times for all classes. As we can see, the probability of an empty queue for Class 1 grows to almost 25% by the end of the time horizon. This indicates that
Figure A.1: Simulation results for a two-class base case with no abandonment: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ (middle panel); and (iii) time-varying staffing level (bottom panel), with $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 1000 independent runs.
Figure A.2: Simulation comparison for a two-class base case with no abandonment: (i) arrival rates (top panel); (ii) simulated class-dependent delay $\mathbb{E}[w_i]$ (top middle panel); (iii) simulated class queue length $\mathbb{E}[Q_i]$ (bottom middle panel); and (iv) simulated probability of empty class queue $P(Q_i(t) = 0)$ (bottom panel), with $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 1000 independent runs.
Figure A.3: Simulation results for a two-class base case with no abandonment and $n = 200$: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $\mathbb{P}(V_i(t) > w_i)$ (middle panel); and (iii) time-varying staffing level (bottom panel), with $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 1000 independent runs.

the system is no longer operating in an efficiency driven regime with respect to Class 1 and violates the assumptions used in the derivation of our control policies. For a finite scale, we conclude that the stabilization effort of the second order TPoD information comes at the expense of the first order PWT and queue length stability. After sufficient time, the nature of our control policy is such that our model is forced into a regime in which our underlying assumptions about model structure begin to break down, resulting in decreased effectiveness of all stabilizing efforts. Although a potential limitation, we remark that these difficulties are mitigated as the size of the system grows larger. We give the base case under a system size where $n = 200$ in Figures A.3 and A.4. In the given time-horizon, we observe the larger scale both stabilizing the TPoD with less periodicity and significantly reducing the variance in queue length.
Figure A.4: Simulation comparison for a two-class base case with no abandonment and $n = 200$: (i) arrival rates (top panel); (ii) simulated class-dependent delay $E[w_i]$ (top middle panel); (iii) simulated class queue length $E[Q_i]$ (bottom middle panel); and (iv) simulated probability of empty class queue $P(Q_i(t) = 0)$ (bottom panel), with $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 1000 independent runs.
Stabilizing Expected Customer Delay

In the previous example, we saw that while the control policies proposed can stabilize the TPoD at a moderate system size, the expected PWT is not properly stabilized. If a system manager instead decides that achieving desired PWT is of greater importance, we show that setting $\alpha_i = \alpha = 0.5$ can be useful in implementation of staffing and scheduling decisions, just as in the main work. By setting these alpha values, we remove both the second order staffing term $c(t)$ and the prioritization regulators, $k_i(t)$ for each class. Without the regulators, the system will schedule customers into service solely based on their current delay in the system. This first order term will no longer be overcome as $t$ grows to be large, and the system will not create an absolute priority class as in the previous example. Simulation experimentation was conducted to verify this result and the output is given in Figure A.5. We see that customer delay remains constant throughout the time horizon and the queue length neither grows or decreases significantly. However, the bottom panel shows that the probability of empty queues increases over time. Unlike the base case, both queues experience approximately equivalent increases simultaneously. Even though the expected queue lengths do not approach empty, the system still experiences increasing variability in the frontier process $\hat{H}$. From Corollary A.1.2 we know that the random terms in the expression for $\hat{H}$ are control independent. Thus, the increasing volatility of the system over time will occur regardless of the prioritization regulators chosen. As a result, the queue lengths becomes more variable and the probability of zero queue length increases (larger queue lengths will also become more likely ensuring the average remains the same).

Constant Arrivals

To support Corollary A.1.6, we provide a constant arrival case. Figure A.6 show the effective TPoD stabilization. We show in Figure A.7 the control functions of the system. As discussed after the corollary, the prioritization regulators grow at different rates and Class 1 increasingly dominates Class 2. As $t$ grows large, the $\kappa_1$ and $\kappa_2$ approach positive and negative infinity, respectively, on the order of $O(\sqrt{t})$. For a fixed system scale ($n$), the prioritization regulator will eventually overcome the scaling of $1/\sqrt{n}$ to have a greater impact on scheduling decision than the HOL wait times. In addition, the safety stock $c(t)$ approaches zero as expected. The total staffing simply becomes the summation of class-specific staffing to match input rate with output (i.e. $\lambda_i = m_i \cdot \mu_i$). From the bottom panel of the control figure, we show the analytical expression for $\sqrt{\text{Var}[\hat{H}(t)]}$ compared to the simulated estimation of the standard deviation of frontier process. As the expected Class 1 queue length decreases to zero over time, the proposed analytical expression for variance becomes less accurate at modeling the true system frontier process variance. The simulated values appear to stabilize over time in contrast to our analytical curve. Similar to previous discussions, this effect is most likely due to the frequent queue depletion of Class 1. However, once the queue length reaches zero it can no longer decrease further. Because we have a staffing function that approaches matching total service capacity with total demand, the expected queue
Figure A.5: Simulation comparison for a two-class base case with no abandonment: (i) arrival rates (top panel); (ii) simulated class-dependent delay $\mathbb{E}[w_i]$ (top middle panel); (iii) simulated class queue length $\mathbb{E}[Q_i]$ (bottom middle panel); and (iv) simulated probability of empty class queue $\mathbb{P}(Q_i(t) = 0)$ (bottom panel), with $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.5$, $\alpha_2 = 0.5$, and 1000 independent runs.
Figure A.6: Simulation results for a two-class base case with no abandonment: (i) arrival rates (top panel); (ii) simulated class-dependent TPoD $\mathbb{P}(V(t) > w_i)$ (middle panel); and (iii) time-varying staffing level (bottom panel), with $w_1 = 0.5$, $w_2 = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, and 1000 independent runs.

lengths will eventually stabilize. As a result, the frontier process begin to stabilize and deviate from our proposed values.

A.1.8 Conclusions

Without the inclusion of abandonment, the multiclass system becomes at once less stochastic but more variable. The mathematical results developed in this Appendix show that as scale increases, we can achieve asymptotic stability, just as is in the main work. In contrast, the variance of the frontier process and the resulting scheduling prioritization regulators both now increase in magnitude with time. For an infinite $n$ and finite $t$, the desired quality of service was shown to be reached. However, in simulation experimentation of finite $n$, we showed that our proposed control functions have unintended effects
Figure A.7: Computed control functions for a two-class base-case with no abandonment and constant arrival rates: \( m(t), c(t), \kappa_i(t) \) and \( \sigma(t), i = 1, 2 \).
on the system. The increasing magnitude of regulators in time results in a class becoming absolutely prioritized (i.e. always routed into service next if waiting). Although initially effective at stabilizing the TPoD, this absolute scheduling forces the system into a state where expected queue length tends to zero. This violates an underlying assumption of positive queue lengths and wait times for all classes presumed in the mathematical development of our control. As a result, the effectiveness of our method at stabilizing TPoD at finite \( n \) will break down at sufficiently large \( t \). For practical implementation, we submit that the PWT is a better QoS metric to focus on system stabilization. We were able to show that by controlling the PWT directly, the resulting queue lengths and system behavior remained stable (i.e. no class becoming absolutely prioritized). Thus, for the case of no abandonment, customer delay should be controlled to achieve overall stable results in time for a finite system.
Appendix B

Appendix for Chapter 3

In this appendix we provide a proof for the FWLLN of the base BSS model under the special initial condition of all initial battery charges and ages set to be zero. In addition, we give the equations for fluid model quality stabilization in the presence of residual battery charge.

B.1 Proof of Theorem 3.3.1

We consider a sequence of queues, which is indexed by the number of batteries in the system $n$ as given in §3.2.2. We consider a finite time interval $[0, T]$. We develop an alternative expression for the exiting process in terms of HOL age and initial battery content as

$$E_n(t) = N_n(t - W_n(t)) + (n - B_{0,n}(t))$$

(B.1)

$$= n^{1/2} \hat{N}_n(t - W_n(t)) + n\Lambda(t - W_n(t)) + (n - B_{0,n}(t))$$

(B.2)

$$= E_{1,n}(t) + E_{2,n}(t) + E_{0,n}(t),$$

(B.3)

where the first (second) term of the first equality counts the number of new (initial) batteries that have exited the system by time $t$ and the second equality follows by the alternative representation of the arrival process give in (3.4) and where

$$E_{1,n}(t) \equiv n^{1/2} \hat{N}_n(t - W_n(t))$$

(B.4)

$$E_{2,n}(t) \equiv n\Lambda(t - W_n(t))$$

(B.5)

$$E_{0,n}(t) \equiv n - B_{0,n}(t).$$

(B.6)
B.1.1 FWLLN for $W_n$

We first establish a FWLLN for $W_n$ following the compactness approach given in [5, 63]. We will show (i) the sequence $W_n$ is C-tight, which implies that every subsequence has a convergent subsequence with a limit in $C$; and (ii) every convergent subsequence converges to the same limit.

**Tightness of $\{W_n\}$**

To establish tightness we develop bounds for both the stochastic variation and modulus of continuity.

**Stochastic boundedness.** The sequence $\{W_n\}$ is bounded because $W_n(t) \geq 0$ and $W_n(t)$ increases at most at rate 1. In a finite interval of length $T$ with a finite initial age process, the upperbound on $W_n(t)$ can be given as $W_n(t) \leq T$. Thus, we have $\lim_{n \to \infty} W_n(t) \leq T$ where $t \in [0, T]$

**Modulus of continuity.** We have the modulus of continuity is bounded above by the rate of 1, $W_n(t + \delta) - W_n(t) \leq \delta$ for $\delta \geq 0$ and $t \in [0, T]$. To bound $W_n(t) - W_n(t + \delta)$, we define the second exiting term process across the interval $[t, t + \delta]$ as

$$\bar{E}_{2,n}(t, \delta) \equiv \bar{E}_{2,n}(t + \delta) - \bar{E}_{2,n}(t) = \int_{t-W_n(t)}^{t-W_n(t)+\delta} \lambda(s) ds \tag{B.7}$$

The integrand in (B.7) is bounded below by the constant $\lambda^\dagger$, which yields a lower bound on $E_{2,n}(t)$,

$$\bar{E}_{2,n}(t, \delta) \geq \int_{t-W_n(t)}^{t-W_n(t)+\delta} \lambda^\dagger ds = (W_n(t + \delta) - W_n(t) + \delta) \lambda^\dagger, \tag{B.8}$$

and an upperbound on the HOL age difference

$$W_n(t + \delta) - W_n(t) \leq \frac{\bar{E}_{2,n}(t, \delta)}{\lambda^\dagger}. \tag{B.9}$$

From the convergence shown in (3.3) and (3.14) and the relation given in (3.4), the second exiting term converges as,

$$\bar{E}_{2,n}(t) \Rightarrow \bar{E}_2(t) = (\Lambda(t) - 1)^+ \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty. \tag{B.10}$$

Thus, the limiting differential exiting term $\bar{E}_{2,n}(t, \delta)$ is bounded above by the maximum arrival rate integrated over $\delta$ time units, the constant $\delta \lambda^\dagger$. Therefore, the $\limsup_{n \to \infty} \{W_n(t + \delta) - W_n(t)\} \leq \delta \frac{\lambda^\dagger}{\lambda^\tau}$ so that the modulus of continuity is bounded as

$$\limsup_{n \to \infty} \{|W_n(t) - W_n(t - \delta)|\} \leq \delta \max\left\{\frac{\lambda^\dagger}{\lambda^\tau}, 1\right\} \tag{B.11}$$

Hence, $W_n$ is tight. In addition, (B.11) also implies that the limit of every convergent subsequence of
$W_n$ is in $C$ and is Lipschitz continuous, where $C$ is the subset of $D$ that is restricted to continuous functions.

**Limit of Convergent Subsequence of $\{W_n\}$**

Let $W_{n_k}$ be a convergent subsequence with the limit $w^*$ (i.e. $W_{n_k} \Rightarrow w^*$). (3.4) implies the convergence along the subsequence $W_{n_k}$ to be

$$\tilde{E}_n(t) \Rightarrow E^*(t) \equiv \Lambda(t - w^*(t)) + 1 - B_0(t) = E(t) = \Lambda(t). \quad (B.12)$$

Taking the derivative of (B.12) yields

$$\lambda(t) = \lambda(t - w^*(t))(1 - \dot{w}^*(t)) + \lambda(t)1\{\Lambda(t) > 1\} \quad (B.13)$$

implying

$$\dot{w}(t) = \frac{\lambda(t)}{\lambda(t - w^*(t))}1\{\Lambda(t) > 1\} + 1\{\Lambda(t) < 1\} \quad (B.14)$$

which coincides with the ODE give in Theorem 3.3.1.

**B.1.2 FWLLN for Key Performance Indicators**

We derive the FWLLN for key performance indicators utilizing the convergence of HOL age and continuous mapping theorem.

**HCL.** The HCL process can be directly translated into age by the charge function,

$$L^\uparrow_n(t) = C(W_n(t), t) \Rightarrow C(w(t), t) = L^\uparrow(t), \quad (B.15)$$

where convergence holds by the continuous mapping theorem.
The scaled total charge level is given as

$$\bar{Y}_n(t) = \frac{1}{n} \sum_{i=N_n(t-W_n(t))}^{N_n(t)} \min\{t - \eta_i, 1\} + t \cdot \bar{B}_{0,n}(t)$$  

(B.16)

$$= \int_{t-W_n(t)}^{t} \min\{t - s, 1\} \lambda(s) ds + t \cdot \bar{B}_{0,n}(t)$$  

(B.17)

$$\Rightarrow \int_{t-w(t)}^{t} \min\{t - s, 1\} \lambda(s) ds + t \left(1 - \int_{0}^{t} \lambda(s) ds\right)^+$$  

(B.18)

**Charge distribution.**

$$\bar{B}_n(t,y) = \frac{1}{n} \sum_{i=N_n(t-W_n(t))}^{N_n(t)} 1\{t - \eta_i < y\} + \bar{B}_{0,n}(t) \cdot 1\{t < y\}$$  

(B.19)

$$= \int_{t-W_n(t)}^{t} 1\{t - s < y\} \lambda(s) ds + \bar{B}_{0,n}(t) 1\{t - \eta_i < y\}$$  

(B.20)

$$\Rightarrow \int_{t-w(t)}^{t} 1\{t - s < y\} \lambda(s) ds + \left(1 - \int_{0}^{t} \lambda(s) ds\right)^+ 1\{t < y\}$$  

(B.21)

$$= \int_{t-(w(t) \wedge y)}^{t} \lambda(s) ds + \left(1 - \int_{0}^{t} \lambda(s) ds\right)^+ 1\{t < y\}$$  

(B.22)

We have shown the FWLLN for key performance indicators of the base BSS model under the special initial condition of all batteries having zero charge.

### B.2 Proposed Fluid Model Functions for Stabilization of FHCL with Residual Charge

Although we were not able to develop an effective fluid approximation algorithm for service quality control in the case of residual charge, we present the corresponding fluid equations that would be necessary to create an algorithm in the same manner as Algorithm 1. We give the ODE for both the HOL age and the resulting necessary charge rate function ODE for stabilizing HCL.

$$\dot{w}(t, \mu) = 1 - \frac{\lambda(t)}{\int_{0}^{w(t,\mu)} \lambda(t-s)f\left(\int_{t-w(t,\mu)+s}^{t} \mu(x) dx\right) ds} \quad t > t_D,$$  

(B.23)

$$\mu(t) = \mu(t-w(t,\mu)) \frac{\lambda(t)}{\int_{0}^{w(t,\mu)} \lambda(t-s)f\left(\int_{t-w(t,\mu)+s}^{t} \mu(x) dx\right) ds} \quad t > t_D.$$  

(B.24)
In the algorithm presented in the main work, we were able to solve first for HOL age and then use the resulting function to solve for required charge. In this case, the residual charge affects the HOL age and we must solve simultaneously for the charge rate function and HOL age.