Dynamic Control of Multiclass Multi-Episode Service Systems with Waiting Time Dependent Transitions

by

Amirmohammad Marshalpirgheybi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Management Sciences

Waterloo, Ontario, Canada, 2026

© Amirmohammad Marshalpirgheybi 2026

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis is based on the manuscript entitled "Dynamic Control of Multiclass Multi-Episode Service Systems with Waiting Time Dependent Transitions," co-authored with Hossein Abouee-Mehrizi, Nasser Barjesteh, and Brendan Wylie-Toal. I was the primary contributor to the mathematical formulation, heavy-traffic analysis, policy development, computational implementation, numerical experiments, and preparation of the manuscript. The co-authors contributed to the conceptual development of the research, methodological refinement, interpretation of the results, and revision of the manuscript. The thesis was prepared by me under the supervision of Professors Hossein Abouee-Mehrizi and Nasser Barjesteh.

Abstract

Motivated by rehabilitation services, this thesis studies a service system with multiple heterogeneous customer classes and servers in which customers may require multiple episodes of service to achieve a desired outcome. After each episode, the service outcome is observed, and customers who do not achieve the desired outcome transition to another class for additional service. Empirical evidence indicates that prolonged waiting can worsen service outcomes and increase future congestion. The model therefore allows service outcomes and class transitions to depend jointly on the service rate employed and the waiting time experienced.

The system manager makes joint dynamic scheduling, service-rate, and customer-rejection decisions to minimize a long-run average expected cost comprising waiting, service, abandonment, rejection, and outcome-related transition costs, subject to firm throughput-time deadlines. Because the resulting stochastic control problem is analytically intractable, the thesis develops a Brownian approximation in a heavy-traffic regime. Under complete resource pooling, the Brownian control problem is reduced to a one-dimensional workload formulation. Its optimal solution is characterized as a two-sided barrier policy with a state-dependent drift rate and workload configuration.

The workload solution is translated into a tractable control policy for the original service system. A comprehensive simulation study shows that the proposed policy substantially outperforms static and myopic benchmarks. The results demonstrate that explicitly accounting for waiting-time-dependent outcomes is essential: policies that ignore these effects can perform poorly because they optimize immediate costs without accounting for the effect of current waiting on future congestion.

The analysis yields several structural insights. Scheduling and service-rate control affect congestion through distinct mechanisms and perform best when used jointly. Scheduling priorities balance immediate costs against downstream congestion effects, shifting toward classes whose waiting times most strongly amplify future congestion as the system becomes more congested. Service-rate control exhibits a regime-switching structure: under low congestion it emphasizes service outcomes, whereas under high congestion it emphasizes congestion mitigation. Finally, although abandonment is costly, it can partially offset the adverse congestion effects of waiting-time-dependent transitions and therefore acts as a partial substitute for dynamic service-rate control.

Acknowledgements

I am deeply grateful to my supervisors, Professors Hossein Abouee-Mehrizi and Nasser Barjesteh, for their guidance, encouragement, and thoughtful feedback throughout this research.

Above all, I would like to express my deepest gratitude to my girlfriend, Melika, the love of my life. Her unwavering love, patience, encouragement, and belief in me sustained me throughout every stage of this journey. She stood beside me during the most demanding moments, celebrated every achievement with me, and continually reminded me of what matters most. Her presence brought strength, peace, and happiness to my life, and this thesis would not have been the same without her constant support. I am profoundly grateful to share my life with her, and words cannot fully express how much she means to me.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

List of Figures ix

List of Tables xi

List of Abbreviations xiii

List of Symbols xiv

1 Introduction 1

2 Literature Review 6

3 Model and Problem Formulation 11

3.1 The Pediatric Rehabilitation Setting 11

3.2 A Class of Multi-Episode Service Systems 12

3.3 A Stochastic Processing Network Formulation 13

3.3.1 Control Policy. 14

3.3.2 System Dynamics. 15

3.3.3 Stochastic Control Problem. 17

4 Approximating Brownian Control Problem **19**

4.1 Heavy-Traffic Assumption and Asymptotic Regime 19

4.2 Brownian Control Problem 23

5 Equivalent Workload Formulation **26**

6 Solution to the Equivalent Workload Formulation **31**

6.1 Barrier Policies 31

6.2 Bellman Equation 32

6.3 Solution to the Bellman Equation and the Equivalent Workload Formulation 34

6.4 First-Stage Problem 36

7 Proposed Policy **38**

8 Simulation Study **41**

8.1 System Parameters in the Base Case 41

8.2 Policy Parameters and the Structure of the Proposed Policy 43

8.3 Description of Policies 47

8.4 Results and Managerial Insights 48

8.5 A Discrete, Non-Preemptive Implementation 51

9 Conclusion **53**

References **55**

Appendix A Supplementary Material for the Model **62**

A.1 Characterization of the Cumulative Routing Processes 62

A.2 An Alternative Characterization of the Cumulative Routing Processes 62

Appendix B Formal Derivation of the Approximating Brownian System **65**

B.1 Arrival, Abandonment, Service, Routing, and Flow Processes. 66

B.2 Scaled Queue Length Process. 71

B.3 Scaled Age Process. 72

B.4 Centered and Scaled Cumulative Cost Process. 73

B.5 Approximating Brownian System 74

Appendix C Proofs of the Technical Results **76**

C.1 An Initial Value Problem Related to the Bellman Equation 76

C.2 Miscellaneous Proofs 78

C.3 Auxiliary Results 89

Appendix D Supplementary Material for the Simulation Study **92**

List of Figures

3.1 Patient flow in a pediatric speech-language therapy program. 11

3.2 Our stochastic processing network with $I = 5$ customer classes and $K = 2$ servers. Arrows from buffers to servers represent activities: solid arrows denote basic activities, and dotted arrows denote nonbasic activities. Upward arrows represent abandonment. Customers may receive up to two episodes. Horizontal arrows represent possible post-service transitions: Class 1 customers may transition to class 5, while class 2 customers may transition to classes 3 or 4. 14

8.1 Transition probability kernels for low-type (left panel) and high-type (right panel) customers as functions of the service rate. Higher service rates and longer waiting times increase the likelihood of requiring additional service. 42

8.2 Effective state cost per unit of work (effective cost) and optimal workload configuration for each class, as a function of the workload, in the base case ($\Delta_0 = 0$). The upper barrier is $u^\star = 7.91$. 45

8.3 Effective state cost per unit of work (effective cost) and optimal workload configuration for each class, as a function of the workload, in the base case ($\Delta_0 = 0$) when $c_{kj} = 0$ for $k = 1, 2$ and $j \in \mathcal{J}$ (i.e., waiting times do not affect service outcomes and transitions). The upper barrier is $u^\star = 16.92$. 45

D.1 Service cost function of activities $j = 1, \ldots, 4$ evaluated at a waiting time of zero for the numerical example presented in Section 8. 93

D.2 The optimal value function $v^\star$ of the workload formulation for the base case of the numerical example presented in Section 8. 93

D.3 Effective state cost per unit of work and optimal workload configuration, as a function of workload, with $\Delta_0 = 0.11$ and $\Delta = 0.0$; all other parameters are as in the base cas. 94

D.4 Effective state cost per unit of work and optimal workload configuration, as a function of workload, with $\Delta_0 = 0.03$ and $\Delta = 0.32$; all other parameters are as in the base case. 95
D.5 Effective state cost per unit of work and optimal workload configuration, as a function of workload, with the sensitivity of service outcomes and transitions to waiting times halved; all other parameters are as in the base case. 95
D.6 Effective state cost per unit of work and optimal workload configuration, as a function of workload, with the sensitivity of service outcomes and transitions to waiting times doubled; other parameters are as in the base case. 96
D.7 Effective state cost per unit of work and optimal workload configuration, as a function of workload, with abandonment rate $\ell_i = 0.03$ for $i \in \mathcal{I}$; all other parameters are as in the base case. 96
D.8 Effective state cost per unit of work and optimal workload configuration, as a function of workload, with abandonment rate $\ell_i = 0.1$ for $i \in \mathcal{I}$; all other parameters are as in the base case. 97
D.9 Effective state cost per unit of work and optimal workload configuration, as a function of workload. The abandonment rates are $\ell_i = 0.03$ for $i \in \mathcal{I}$ and abandonment costs are halved; other parameters are as in the base case. 97
D.10 Effective state cost per unit of work and optimal workload configuration. The Abandonment rates are $\ell_i = 0.03$ for $i \in \mathcal{I}$ and abandonment costs are doubled; other parameters are as in the base case. 98
D.11 Effective state cost per unit of work and optimal workload configuration, as a function of workload, with waiting costs halved; all other parameters are as in the base case. 98
D.12 Effective state cost per unit of work and optimal workload configuration, as a function of workload, with waiting cost doubled; all other parameters are as in the base case. 99

List of Tables

8.1 Nominal scheduling and service rate vectors, and service costs at the nominal service rates and zero waiting time for the base case. Service rates of nonbasic activities are optimized as described in Section 6.4. 43

8.2 Summary of the six policies considered in the simulation study. 47

8.3 Comparative evaluation of the long-run average cost of the various policies. 48

8.4 Comparative evaluation of the long-run average costs in the discrete, non-preemptive setting. 51

D.1 Parameters of the transition kernels and the probability that a customer does not reach the desired service outcome after an episode of service. Entries marked "–" denote parameters that are not applicable. 92

D.2 Parameters associated with buffers. 93

D.3 Parameters associated with pushing controls. 93

D.4 Comparative evaluation of the long-run average cost of the various policies for three different deadlines (δ_i for $i \in \mathcal{I}$). Deadlines of all classes are the same. 100

D.5 Comparative evaluation of long-run average costs, varying only the sensitivity of service outcomes and transitions to waiting times (c_{1j} and c_{2j} for $j \in \mathcal{J}$). 100

D.6 Comparative evaluation of long-run average costs, varying only the abandonment rate (ℓ_i for $i \in \mathcal{I}$). 101

D.7 Comparative evaluation of long-run average costs, varying only the abandonment costs (π_i for $i \in \mathcal{I}$) at the abandonment rate of $\ell_i = 0.03$ for $i \in \mathcal{I}$. 102

D.8 Comparative evaluation of long-run average costs, varying only the waiting costs (c_{3j} for $j \in \mathcal{J}$). 103

D.9 Optimized baseline service rates (Δ_0) the S-OS and D-DO policies in Table D.4. . . 103

D.10 Optimized baseline service rates (Δ_0) of the S-OS and D-DO policies in Table D.5. 104

D.11 Optimized baseline service rates (Δ_0) of the S-OS and D-DO policies in Table D.6. 104

D.12 Optimized baseline service rates (Δ_0) of the S-OS and D-DO policies in Table D.7. 105
D.13 Optimized baseline service rates (Δ_0) of the S-OS and D-DO policies in Table D.8. 105
D.14 Comparative evaluation of the long-run average cost of the various policies for three different deadlines (δ_i for $i \in \mathcal{I}$) in the discrete non-preemptive setting of Section 8.5. 106
D.15 Comparative evaluation of long-run average costs, varying only the sensitivity of service outcomes and transitions to waiting times (c_{1j} and c_{2j} for $j \in \mathcal{J}$) in the discrete non-preemptive setting of Section 8.5. 106
D.16 Comparative evaluation of long-run average costs, varying only the abandonment rate (ℓ_i for $i \in \mathcal{I}$) in the discrete non-preemptive setting of Section 8.5. 107
D.17 Optimized baseline service rates (Δ_0) of the S-OS and D-DO policies in Table D.14. 108
D.18 Optimized baseline service rates (Δ_0) of the S-OS and D-DO policies in Table D.15. 108
D.19 Optimized baseline service rates (Δ_0) of the S-OS and D-DO policies in Table D.16. 108

List of Abbreviations

BCP	Brownian control problem
CRP	Complete resource pooling
D-DO	Dynamic scheduling with dynamic service rates and an optimized baseline
D-DZ	Dynamic scheduling with dynamic service rates and a zero baseline
ED	Emergency department; where noted, efficiency-driven regime
EWF	Equivalent workload formulation
FCFS	First-come, first-served
ICU	Intensive care unit
LOS	Length of stay
QED	Quality-and-efficiency-driven regime
SPN	Stochastic processing network
SPP	Static planning problem

List of Symbols

$\mathcal{I}$	Set of customer classes (buffers)
$\mathcal{J}$	Set of service activities
$\mathcal{K}$	Set of servers or resources
$Q(t)$	Queue-length vector at time t
$n(t)$	Scheduling or activity-allocation vector at time t
$\mu(t)$	Service-rate vector at time t
$W(t)$	Aggregate workload process at time t
$\eta(t)$	Diffusion-scale service-rate adjustment
$Y(t)$	Brownian scheduling/idleness control process
$G(t)$	Brownian rejection control process
Λ	Mapping from queue lengths to customer ages
δ	Vector of throughput-time deadlines
κ_L, κ_U	Lower- and upper-barrier pushing costs

Chapter 1

Introduction

In many service systems, waiting affects more than the timing of service: it can increase the amount of service customers ultimately require and reduce the likelihood of achieving a desirable outcome. Empirical evidence documents such effects in emergency departments (Huang et al. 2010, Su et al. 2024, Liu et al. 2025), intensive care units (Chan et al. 2017), rehabilitation services (Abouee-Mehrizi et al. 2026), call centers (Reich et al. 2010), and restaurants (De Vries et al. 2018). These findings imply that congestion may be self-reinforcing: longer waits worsen service outcomes or increase service requirements, thereby generating additional future demand and further congestion.

The theoretical literature on waiting-time-dependent service requirements and outcomes has largely concentrated on single-episode systems, in which each customer receives one episode of service and then departs. In these models, waiting-time effects are commonly represented by service rates that decline exogenously with waiting time (Dong et al. 2015, 2025). This representation does not capture an important feature of many service systems: customers may require several episodes of service, and the need for future service is revealed only after the outcome of the current episode is observed.

Pediatric rehabilitation services, including speech-language therapy, provide a canonical example of such a multi-episode system (Lurie Children's Hospital 2023). A customer may receive a sequence of treatment episodes separated by waiting periods before achieving the desired outcome. The total amount of service required is not known at entry. Instead, it is revealed progressively: after each episode, the service outcome is assessed and a decision is made regarding whether additional service is required (Cunningham et al. 2017). Moreover, the duration of an episode—or, equivalently, its service rate—is itself a managerial decision (Harding et al. 2024). A customer who does not achieve the desired outcome may return for another episode, possibly in a different class that reflects an updated clinical or operational status. The probabilities of these post-service transitions depend

jointly on the service rate used during the episode and the waiting time experienced before service. Consequently, waiting-time-dependent transitions directly link current congestion to future demand.

These systems create a dynamic control problem in which the manager must determine which customers to serve, how intensively to serve them, and, when necessary, whether to remove customers from the system. Scheduling and service-rate decisions affect immediate costs, but they also alter future congestion by changing service outcomes and transition probabilities. A decision that appears attractive myopically may increase waiting for other classes, worsen their outcomes, and create additional downstream service demand. Therefore, policies based solely on immediate cost reduction can perform poorly. This thesis develops an analytical framework for the joint dynamic control of multiclass, multiserver, multi-episode service systems with waiting-time-dependent outcomes and transitions.

Multi-episode service systems also possess operational features that require a richer modeling framework. In pediatric rehabilitation, clinicians typically maintain concurrent caseloads, and each patient occupies a recurring appointment slot during an episode; hence, each patient consumes a fraction of a clinician's capacity over an extended period (Children's Hospital of Richmond 2023). In addition, publicly funded services are often available only during a limited eligibility window. This restriction imposes a firm deadline on a customer's total throughput, or sojourn, time: a child who has not achieved the desired outcome before becoming ineligible must leave the program (McGill et al. 2020, KidsAbility 2026). Customers may also abandon voluntarily when waits become excessive (McGill et al. 2020). These mechanisms introduce state constraints and alternative channels through which congestion is reduced, and they interact nontrivially with waiting-time-dependent outcomes and transitions.

Motivated by these features, this thesis formulates a stochastic processing network (SPN) for multiclass, multiserver, multi-episode service systems. Customers may require several episodes before achieving their desired outcome. After each episode, the outcome is observed; customers who require additional service transition to another class, possibly retaining or changing their status. Both the service outcome and the subsequent class transition depend on the episode-specific service rate and on the waiting time preceding that episode.

Within this framework, the system manager dynamically chooses (i) the customers assigned to each server, (ii) the service rate used for each episode, and (iii) whether customers should be rejected from the system. The objective is to minimize the long-run average expected cost. The cost structure includes service costs that depend on both waiting time and service rate, abandonment costs incurred when impatient customers leave before service, and rejection costs incurred when the manager removes customers. Customers are also subject to firm throughput-time deadlines; reaching a deadline forces the customer's removal from the system.

The exact joint scheduling, service-rate, and rejection control problem is analytically and computationally intractable. Following the heavy-traffic methodology of Harrison (1998, 2000, 2003), this thesis approximates the original SPN by a Brownian control problem in a regime where system capacity is nearly fully utilized. Under complete resource pooling, the servers' overlapping capabilities allow the network to behave, at the diffusion scale, as a single pooled resource. We show that the Brownian control problem is equivalent to a one-dimensional drift-rate control problem whose state is the aggregate workload.

The resulting drift-rate control problem admits a two-sided barrier policy. The optimal policy keeps workload between endogenously determined lower and upper barriers and selects a state-dependent drift rate and workload configuration within this interval. These controls have direct interpretations in the original service system. The workload configuration induces a scheduling rule, the drift-rate choice determines the service-rate regime, and the reflecting controls correspond to idling, activating nonbasic activities, or rejecting customers when workload reaches a boundary.

The scheduling component first gives priority to customers at risk of violating their throughput-time constraints. It then allocates the remaining capacity to move the queue-length configuration toward the optimizer of the workload formulation. This allocation is governed by effective state costs that combine immediate congestion costs, such as waiting and abandonment costs, with forward-looking terms that quantify how retaining workload in each buffer affects future congestion through abandonment and waiting-time-dependent transitions. Hence, scheduling priorities change with the workload level rather than remaining fixed. The service-rate component also varies with congestion: one service-rate vector is used when workload is low and another when workload is high. At the barriers, additional controls keep the workload within the prescribed interval.

Modeling Contributions. This thesis develops a unified framework for heterogeneous service systems in which customers may require multiple episodes, service outcomes and class transitions are stochastic and jointly dependent on waiting times and service rates, customers may abandon, and total sojourn times are subject to firm deadlines. It jointly incorporates scheduling, endogenous service-rate selection, and rejection control. To the best of our knowledge, this is the first study to analyze these three control dimensions together in a multi-episode SPN with waiting-time-dependent transitions. The framework also yields an implementable dynamic policy rather than only a structural characterization of the limiting control problem.

Technical Contributions. The Brownian approximation leads to a drift-rate control problem with linear drift-control costs and a general convex piecewise-linear state-cost function. The problem contains an uncontrollable state-dependent drift component, endogenously determined reflecting barriers, and a workload configuration that depends jointly on workload and the derivative of the value function. This coupling captures the effects of abandonment and waiting-time-dependent

transitions and makes the optimal policy inherently non-greedy. Despite this dependence, we show that the Bellman equation can be solved efficiently through a one-dimensional search for the root of a strictly increasing function. We also establish the optimality of a two-sided barrier policy with state-dependent drift and workload configuration. Finally, the analysis identifies a role for the service rates of nonbasic activities—activities that are absent from the nominal static plan. Their service rates are not determined by the static planning problem, but they affect the costs of the boundary-pushing controls and therefore influence the optimal workload policy.

Structural Insights. The analysis and simulation study reveal several properties of effective congestion control in multi-episode systems. First, scheduling priorities depend on both immediate and future congestion costs. Under low congestion, classes with large immediate costs receive priority. As congestion grows, priority shifts toward classes whose waiting times have the greatest effect on future congestion. Second, service-rate control has a regime-switching structure. When congestion is low, the policy emphasizes improved service outcomes, even when doing so may increase subsequent workload. When congestion is high, it instead emphasizes congestion mitigation, potentially at the expense of outcomes for customers currently in service. Thus, policies that uniformly reduce service rates under congestion can amplify the feedback between waiting and future demand. The gains from service-rate flexibility are nevertheless subject to diminishing returns.

Third, proactive rejection is used as a last-resort barrier regulator. It becomes worthwhile only under extreme congestion, and rejecting customers from a single class is sufficient to implement the upper-boundary control. Fourth, the policy is driven primarily by aggregate workload, which evolves more slowly than the underlying arrival and service processes. It therefore responds to persistent changes in congestion rather than to every short-run queue-length fluctuation; in large systems, relatively small service-rate adjustments are sufficient.

Scheduling, service-rate control, and abandonment affect congestion through different mechanisms. Service-rate control directly changes workload dynamics, whereas scheduling determines which customers wait and thereby changes future transition behavior. Policies that use both control levers consistently outperform policies that rely on only one. Although abandonment is costly, it can partially offset the positive congestion feedback generated by waiting-time-dependent transitions. In this sense, abandonment acts as a partial substitute for dynamic service-rate control and interacts closely with scheduling.

More broadly, waiting-time-dependent outcomes create a feedback loop in which congestion worsens future outcomes and thereby generates additional congestion. Effective policies must therefore be forward-looking: they may accept a larger immediate cost to prevent a substantially larger downstream cost. The numerical results show that explicitly accounting for this feedback is essential. As the sensitivity of outcomes and transitions to waiting increases, myopic policies

deteriorate rapidly relative to the proposed dynamic policy.

Organization. The rest of this thesis is organized as follows. Chapter 2 reviews the literature. Chapter 3 presents the model and problem formulation. Chapter 4 introduces a Brownian approximation of the original control problem, and Chapter 5 reduces it an equivalent workload formulation. Chapter 6 solves the workload formulation, and Chapter 7 interprets its solution as a dynamic control policy for the original service system. Chapter 8 examines the structure of the proposed policy, evaluates its effectiveness through an extensive simulation study, and offers managerial insights. Several appendices supplement the main body. Appendix A provides supplementary material for the model. Appendix B provides a formal derivation of the Brownian control problem. Appendix C contains the proofs of the main technical results. Finally, Appendix D provides supplementary material for the simulation study.

Chapter 2

Literature Review

This thesis relates to three streams of literature: (i) control of queueing systems and their generalizations, stochastic processing networks, in heavy traffic, (ii) service rate control of queueing systems where customers may receive multiple episodes of service, and (iii) queueing systems in which waiting times impact service requirements and outcomes. The first stream of literature studies the control of queueing and stochastic processing networks; see Stidham (1988, 2002) for surveys. From a methodological perspective, this thesis builds on the literature on the dynamic control of queueing systems in heavy traffic, pioneered by Harrison (1988), and later generalized to stochastic processing networks in Harrison (2000, 2003) and Harrison and Williams (2005, 2007). In this literature, the original queueing system is approximated by a Brownian system, which offers greater analytical tractability. Early examples of this approach include Harrison and Wein (1989), which studies a crisscross network, and Harrison and Wein (1990), which studies a multiclass two-station closed queueing network. Since then, this approach has been applied to a wide range of queueing problems; see, e.g., Wein (1992), Kumar (2000), Plambeck (2004), Çelik and Maglaras (2008), Ata and Olsen (2009), Kim et al. (2018), Ata et al. (2020), and Sun and Zhu (2024). Our work is most closely related to studies of stochastic processing networks under complete resource pooling, see e.g., Harrison and López (1999), Ata and Kumar (2005), Dai and Lin (2008), Rubino and Ata (2009), and Ghamami and Ward (2013).

In the heavy-traffic literature, service rate control problems are typically approximated by drift rate control problems for the associated Brownian system. Early examples include Krichagina and Taksar (1992) and Krichagina et al. (1993) which establish the optimality of threshold-type policies. A seminal paper in the drift rate control literature is Ata et al. (2005), which studies a one-dimensional drift rate control problem on a bounded interval. Ghosh and Weerasinghe (2007) extends Ata et al. (2005) by introducing nondecreasing holding costs and an endogenously determined upper barrier. Following this approach, Ata et al. (2019) approximates a gleaning operation with a drift rate control

problem and derives a nested threshold-type staffing policy. Ata and Barjesteh (2022) approximates the control of a multiclass make-to-stock production system with a drift rate control problem with a V-shaped holding cost and a quadratic cost of drift rate control. Gao and Huang (2023) extends Ata and Barjesteh (2022) by incorporating strictly convex state costs and proving asymptotic optimality. Building on Ata and Barjesteh (2022), Alwan and Barjesteh (2025) approximates the control of a multiclass omnichannel hybrid production system with a drift rate control problem with a specific nonnegative piecewise linear holding cost, a quadratic cost of drift rate control, and a negative state-dependent drift component. Another related paper is Ata et al. (2024b) that approximates a multiclass volunteer engagement problem with a drift rate control problem with no holding costs, a piecewise linear cost of drift rate control, and a negative state-dependent drift component. This thesis extends this literature by approximating the control of a stochastic processing network, where waiting times impact service outcomes and transitions, with a drift rate control problem with a general convex piecewise linear holding cost, a linear cost of drift rate control, and state constraints. Our formulation also accommodates both positive and negative state-dependent drift components, arising from the countervailing influences of abandonments and waiting times, whose magnitude is controlled by the workload configuration, and endogenously determines the lower and upper barriers. In addition, our analysis reveals that in stochastic processing networks, the service rates of nonbasic activities, i.e., those not used in the nominal (static) scheduling policy, play a central role in the workload formulation by impacting the cost of the pushing controls, and thereby the optimal policy. Other related papers in this area include Ghosh and Weerasinghe (2010), Budhiraja et al. (2011), Alwan et al. (2024), and Ata et al. (2024a).

Dynamic control of queueing systems with abandonments in heavy traffic has also received considerable attention. Several papers, including ours, assume exponentially distributed patience times, which leads to a linear drift rate in the Brownian control problem; see, e.g., Ward and Kumar (2008), Rubino and Ata (2009), Ghosh and Weerasinghe (2010), Ghamami and Ward (2013), Ibrahim et al. (2017), and Ata et al. (2024b). For extensions that consider non-exponential patience times, see Reed and Ward (2008), Ibrahim and Whitt (2009), Bassamboo and Randhawa (2015), and Kim et al. (2018). This thesis extends this literature by studying a stochastic processing network and incorporating service rate control decisions and waiting time dependent service outcomes and transitions. We show that although abandonments are costly, they can partially mitigate the adverse effects of prolonged waiting times by generating a drift rate that could counteract the drift rate induced by the impact of waiting times on service outcomes and transitions.

An important consideration in our problem is that customers who do not achieve the desired service outcome before reaching their deadlines must be rejected (i.e., removed from the system). This aspect is reminiscent of lead time constraints in make-to-order queueing systems (Çelik and Maglaras 2008, Rubino and Ata 2009, Afèche 2013); see Kempf et al. (2011) for an overview of this

literature. In the lead time control literature, each job is served at most once and departs the system upon completion. Consequently, the due-date constraints translate, in the heavy-traffic regime, into queue length bounds, as discussed in Plambeck (2004). In contrast, in our setting, customers may receive multiple episodes of service, each preceded by a waiting period. Consequently, the system manager must track each customer's age throughout their sojourn in the system, which necessitates careful modeling and imposes nuanced constraints.

The second stream of literature studies service rate control in queueing systems where customers may receive multiple episodes of service. Chan et al. (2012) studies discharge decisions in an ICU setting and provides empirical evidence that simple index policies that account for readmission (return) risk can increase throughput. Chan et al. (2014) studies a single-class queueing system with multiple identical servers where customers may return for additional service. It focuses on threshold-type service rate control policies, which increase the service rate when the queue length exceeds a threshold, and identifies conditions under which such policies improve or degrade performance. Yom-Tov and Chan (2021) extends this literature by incorporating admission control decisions. Shi et al. (2021) develops a readmission prediction model for a hospital and uses it to prescribe daily discharge decisions. This thesis builds on this literature by proposing a joint dynamic scheduling, service rate, and rejection control policy for a multiclass stochastic processing network with heterogeneous servers, where customers are impatient and waiting times impact transition probabilities. In our setting, customers have a deadline on their throughput time, and the control policy explicitly incorporates customer history (e.g., the number of prior service episodes).

The third stream of literature examines queueing systems in which waiting times impact service requirements and outcomes. A growing body of empirical research suggests that prolonged waiting times impact service requirements and outcomes across a variety of settings. For instance, Chan et al. (2017) demonstrates that prolonged waiting times for ICU services are associated with poorer patient outcomes and longer ICU lengths of stay (LOS). Using an instrumental variable design, Su et al. (2024) provide causal evidence that prolonged Emergency Department (ED) boarding leads to longer inpatient LOS, higher likelihood of escalation in required care level, and increased hospital charges. Complementing this, Liu et al. (2025) uses double machine learning methods to establish that the impact of ED boarding on inpatient LOS is heterogeneous across patients, and identifies subgroups for whom prolonged boarding has different effects. Abouee-Mehrizi et al. (2026) identifies analogous dynamics in pediatric rehabilitation, showing that prolonged waiting times are associated with an increase in the number of appointments received before achieving the desired outcome. Other empirical studies include Siegmeth et al. (2005), Chalfin et al. (2007), and Chan et al. (2008) on healthcare delivery, Reich et al. (2010) on call centers, and De Vries et al. (2018) on restaurants.

Several studies have analyzed systems in which waiting times impact service requirements from a theoretical perspective. Much of the early work focused on the scheduling of a fixed number of jobs (with no arrivals) on a single server. For example, Browne and Yechiali (1990) considers a system in which service times increase linearly with waiting times and shows that prioritizing jobs by the ratio of expected initial service time to deterioration rate minimizes makespan. Subsequent work extends this line of research by examining alternative slowdown formulations (e.g., piecewise linear and convex) and alternative objectives (e.g., flow time and maximum tardiness); see Alidaee and Womer (1999) and Cheng et al. (2004) for reviews.

One of the early papers studying a system with arrivals is Posner (1973), which studies a single-class single-server queueing system in which service times depend on waiting times through a step-function and derives the stationary waiting time distribution for the FCFS policy. Subsequent work has extended the analysis to systems with two servers (Brill and Posner 1981) and single-server systems with general inter-arrival and service time distributions (Whitt 1990, Boxma and Vlasiou 2007, Bekker et al. 2009, Palmowski and Vlasiou 2011). Mandelbaum and Pats (1998) extends this literature to a network of Markovian single-class single-server queueing system with state-dependent arrivals, services, and transitions.

Another line of research considers systems with a single class of customers and multiple identical servers. Weerasinghe (2014) derives fluid and diffusion approximations for such a system with impatient customers. Dong et al. (2015) considers a related system in the Quality-and-Efficiency-Driven (QED) regime, and shows that when service times are highly sensitive to system load, the system alternates between QED and ED equilibria, a phenomenon referred to as bistability. Wu et al. (2019) examines the dependence between customer patience times and service requirements, and shows that even moderate dependence can substantially impact performance. Building on this, Wu et al. (2022) considers both exogenous dependence, in which customer patience times and service requirements are stochastically dependent, and endogenous dependence, in which service requirements may evolve while customers wait, and shows that the latter generalizes the former. Further related work includes Selen et al. (2016) and D'Auria et al. (2022).

Most prior work studies single-class systems or networks of single-class stations. Notable exceptions include Hu et al. (2022) and Dong et al. (2025). Hu et al. (2022) studies a multiserver queueing system with two customer classes, where customers may transition between classes while waiting, and proposes a modified $c\mu/\theta$ index policy that balances holding costs, service times, abandonments, and transitions. Dong et al. (2025) studies a multiclass queueing system with multiple identical servers in which waiting times increase service requirements. It develops a simulation-based approximate dynamic programming algorithm that yields close-to-optimal scheduling policies and offers structural insights, such as highlighting the trade-off between immediate cost reduction and

the risk of drifting toward undesirable equilibria.

Most existing work focuses on single-episode service systems, in which waiting time effects are represented as service rates that decrease exogenously with waiting times. In contrast, we study a multi-episode service system with stochastic service outcomes. In this setting, service requirements are not known a priori, and customers may require a sequence of service episodes, each preceded with a waiting period, before achieving the desired service outcome and departing the system. In these systems, service rates are decision variables that can be dynamically adjusted, and waiting time effects manifest as transition probabilities that depend jointly on service rates and waiting times. We develop a tractable and well-performing policy that provides insights into how scheduling, service rate, and rejection decisions should respond to congestion. Our analysis highlights a central trade-off: actions that reduce immediate costs can amplify future congestion by worsening service outcomes, while congestion-mitigating actions may incur high immediate costs. This tension shapes the scheduling and service rate decisions. A key implication is a regime-dependent service rate structure. When congestion is low, the policy favors service rates that reduce immediate costs, e.g., by improving service outcomes, even at the expense of higher future congestion. When congestion is high, it switches to a service rate vector that prioritizes congestion mitigation, even at the expense of current service outcomes.

Chapter 3

Model and Problem Formulation

This section develops the model. Section 3.1 describes the pediatric rehabilitation setting that motivates the study and identifies its salient operational features. Section 3.2 abstracts from this setting to describe the broader class of multi-episode service systems we study. Section 3.3 casts this class as a stochastic processing network and formulates the stochastic control problem.

3.1 The Pediatric Rehabilitation Setting

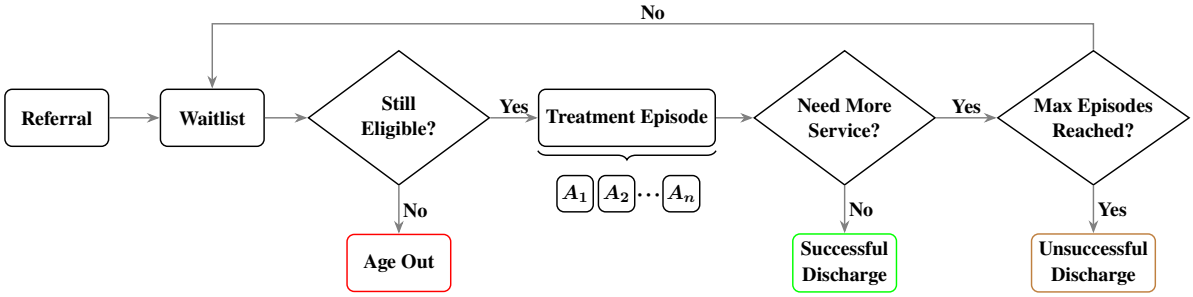


Figure 3.1: Patient flow in a pediatric speech-language therapy program.

A pediatric speech-language therapy program provides communication intervention to young children with language, speech, or fluency disorders. The program is staffed by a team of clinicians, each maintaining a caseload of patients in concurrent treatment. A patient on a clinician's caseload receives appointments on a fixed schedule, typically weekly or biweekly. Each appointment occupies a fraction of the clinician's weekly capacity, so each patient consumes a fraction of the clinician's capacity continuously over the duration of the treatment episode. When a clinician completes an episode for a patient and removes the patient from the caseload, they add another patient from the waitlist to their caseload.

Treatment is delivered over a sequence of episodes, each consisting of several appointments. The number of appointments within an episode determines its duration and, equivalently, the service rate at which the patient is treated. After each episode, the clinician conducts a reassessment to determine whether the patient has achieved the desired outcome. If so, the patient is discharged; if not, the patient returns to the waitlist for a subsequent episode, possibly with an updated clinical status. The total number of episodes, and hence appointments, a patient ultimately requires is unknown at the outset; it is revealed progressively through the sequence of reassessment outcomes.

The system manager exercises two controls: scheduling, which governs which patients are on each clinician's caseload at any moment in time, and service rate, which determines the number of appointments each patient receives in each episode. Patients are eligible only up to a fixed age, typically when they enter school, after which they must leave the program whether or not the desired outcome has been achieved and seek treatment through another provider. Some families, faced with long waiting times, seek private care or disengage voluntarily. A central feature of these programs is that prolonged waiting times adversely affect treatment outcomes (Abouee-Mehrizi et al. 2026). Patients who wait longer tend to require more appointments and are more likely to need an additional episode. Post-treatment transitions are stochastic outcomes determined by the patient's class, the service rate employed, and the waiting time experienced before the episode. Consequently, the system manager's decisions, namely which patient to add to a caseload next and how many appointments each patient receives in each episode, shape not only the immediate schedule but also the future evolution of congestion.

3.2 A Class of Multi-Episode Service Systems

Pediatric rehabilitation belongs to a broader class of service systems characterized by the following features.

Multi-episode service with stochastic outcomes. The service requirement of a customer is unknown at the outset, and is revealed through a sequence of service episodes. After each episode, a stochastic outcome determines whether the customer departs or requires another episode.

Fractional capacity allocation. Each server serves multiple customers concurrently, with each customer consuming a fraction of the server's capacity per unit time.

Waiting time dependent outcomes. The probability of a successful service outcome depends on both the service rate and the waiting time experienced before service. Prolonged waiting times degrade outcomes and increase the likelihood that additional service episodes will be required.

Deadlines, abandonment, and rejection. Each customer has a firm deadline, after which they are no longer eligible for service. Customers may abandon while waiting, and the system manager may

reject customers.

The objective is to minimize the long-run average cost arising from delay, transition (i.e., failure to achieve the desired outcome), abandonment, and rejection. The pediatric speech-language therapy setting maps directly onto this formulation: clinicians are servers; treatment episodes are service episodes; reassessment outcomes drive class transitions; the number of appointments within an episode determines the service rate; age eligibility imposes deadlines; and prolonged waiting degrades outcomes.

3.3 A Stochastic Processing Network Formulation

We model the class of systems described in Section 3.2 as a stochastic processing network (Harrison 2003). A stochastic processing network generalizes a traditional queueing network and consists of three basic elements: customers, servers, and activities. The notion of customers and servers mirrors that of a queueing network. Customers are classified into distinct classes based on their initial status, immutable characteristics, and number of prior service episodes. Each customer class has a dedicated buffer where the customers of that class wait. The distinctive feature of a stochastic processing network is that an activity can accommodate more complex operational tasks beyond those captured in a traditional queueing network, e.g., a customer may be served through multiple activities, each involving a different server. In our setting, each activity involves a unique server serving a single customer from a unique customer class, and all service times are exponentially distributed.[1] Figure 3.2 provides an illustrative example of our model; see Chapter 8 for details. The modeling components shown in the figure will be introduced and explained in what follows.

Let I denote the number of customer classes, J denote the number of activities, and K denote the number of servers. We let $\mathcal{I} = \{1, \ldots, I\}$, $\mathcal{J} = \{1, \ldots, J\}$, and $\mathcal{K} = \{1, \ldots, K\}$ denote the sets of customer classes, activities, and servers, respectively. The customer class and server associated with activity $j \in \mathcal{J}$ are denoted by $i(j)$ and $k(j)$, respectively, and the set of activities that serve class $i \in \mathcal{I}$ is denoted by $\mathcal{J}_i = \{j \in \mathcal{J} : i(j) = i\}$. Since the number of prior service episodes is encoded in the class definition, each class receives either external arrivals (i.e., new customers) or transitions from another class, but not both; see Figure 3.2. External arrivals to class i occur according to a Poisson process with rate λ_i. We let $E_i(t)$ denote the cumulative number of external arrivals to class $i \in \mathcal{I}$ up to time t.

[1] The Markovian assumption is standard in systems where multiple jobs may be served concurrently from a single buffer (see, e.g., Kim et al. (2018) and Hu et al. (2022)), and it significantly facilitates analytical tractability.

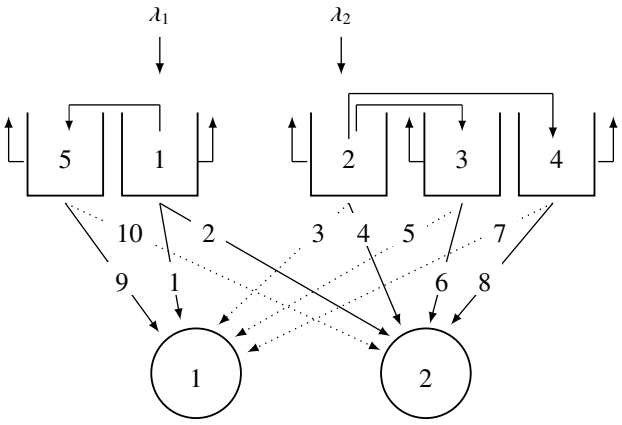


Figure 3.2: Our stochastic processing network with $I = 5$ customer classes and $K = 2$ servers. Arrows from buffers to servers represent activities: solid arrows denote basic activities, and dotted arrows denote nonbasic activities. Upward arrows represent abandonment. Customers may receive up to two episodes. Horizontal arrows represent possible post-service transitions: Class 1 customers may transition to class 5, while class 2 customers may transition to classes 3 or 4.

3.3.1 Control Policy.

The system manager dynamically makes scheduling, service rate, and rejection decisions. The scheduling decision determines the number of concurrent instances of each activity to undertake. Each instance of an activity serves a single distinct customer. Let $n_j(t)$ denote the number of concurrent instances of activity $j \in \mathcal{J}$ undertaken at time t. That is, $n_j(t)$ is the number of class $i(j)$ customers concurrently served through activity j by server $k(j)$ at time t. Letting $n(t) = (n_j(t))$ for $t \geq 0$, we refer to $n = \{n(t) : t \geq 0\}$ as the scheduling process. We focus on head-of-line (within each class) scheduling policies.[2] Letting $A_{kj} \in [0, 1]$ denote the fraction of the capacity of server $k \in \mathcal{K}$ consumed by one instance of activity $j \in \mathcal{J}$, we refer to $A = [A_{kj}]$ as the capacity consumption matrix. In the rehabilitation setting, A_{kj} is the fraction of clinician k's weekly capacity consumed by a patient receiving treatment through activity j.

The system manager sets the service rates by choosing the instantaneous service rate vector $\mu(t) = (\mu_j(t)) \in \mathbb{R}_+^J$, where $\mu_j(t)$ denotes the instantaneous service rate of activity $j \in \mathcal{J}$ at time t. By adjusting the service rates, the system manager adjusts the service durations.[3] We refer to $\mu = \{\mu(t) : t \geq 0\}$ as the service rate process. We denote the set of admissible service rate vectors by $\mathcal{M} \subset \mathbb{R}_+^J$, and the set of admissible service rates for activity $j \in \mathcal{J}$ by $\mathcal{M}_j \subset \mathbb{R}_+$. For analytical tractability, we assume $\mathcal{M}$ is bounded and bounded away from zero. To limit disparities in service rate across customers in the same class, the system manager restricts the variations of each activity's

[2]Customers within each class are ordered by their age, and head-of-line scheduling serves the oldest customers in each class.

[3]In the rehabilitation setting, patients receive treatment at a fixed frequency, e.g., once a week, and the service rate determines the number of appointments within an episode.

service rates over time. We denote the maximum (allowable) variation in the service rate of activity $j \in \mathcal{J}$ by Δ_j.

The system manager may reject customers, i.e., remove them from the system, at any point in time. A rejection policy is represented by an I-dimensional rejection process $O = \{O(t) : t \geq 0\}$, where $O_i(t)$ denotes the cumulative number of class $i \in \mathcal{I}$ customers rejected up to time t.

Accordingly, the control policy is a triple (μ, n, O), where μ is the service rate process, n is the scheduling process, and O is the rejection process. While the model permits preemption and changes to a customer's service rate during an episode, and assumes exponential service times, Section 8.5 shows that our proposed policy remains effective under a non-preemptive implementation with each customer's service rate held constant over each episode, and discrete service times (corresponding to integer number of appointments).

3.3.2 System Dynamics.

Upon completion of a service episode, the customer either departs the system or transitions to another class for further service, provided they have not exceeded the maximum (allowable) number of episodes.[4] Let $\mathcal{B}_i$ denote the set of classes to which a class $i \in \mathcal{I}$ customer may transition upon service completion. Without loss of generality, we take $\mathcal{B}_1, \ldots, \mathcal{B}_I$ to be disjoint. The transition structure is thus acyclic: customers do not revisit classes, and each downstream class (one with no external arrivals) is reachable through a unique transition path from an initial class (one with external arrivals); see Figure 3.2. This structure enables the tractable head-of-line age approximation developed in Chapter 4.

The transition probabilities governing class transitions after each episode depend on both the service rate employed and the waiting time experienced by the customer in the current buffer. Let $w_i(t)$ denote the virtual waiting time in buffer $i \in \mathcal{I}$, that is, the waiting time a hypothetical infinitely patient customer arriving at buffer i at time t would experience (in buffer i). We refer to $w(t) = (w_i(t))$ as the virtual waiting time vector at time t. For each activity $j \in \mathcal{J}$ and class $l \in \mathcal{I}$, the transition probability is specified by a transition kernel $P_{jl} : \mathcal{M}_j \times \mathbb{R}_+ \to [0, 1]$. Specifically, $P_{jl}(\mu_j, w_{i(j)})$ is the probability that a class $i(j)$ customer, served under activity j at service rate μ_j after waiting $w_{i(j)}$ units of time, transitions to class l.

Let $\Phi_l^j(m)$ denote the cumulative decrease in the number of class $l \in \mathcal{I}$ customers as a result of the first $m \in \mathbb{N}$ service completions of activity $j \in \mathcal{J}$; see Appendix A.1 for its construction. Letting $\Phi^j(m) = (\Phi_l^j(m))$, we refer to $\Phi^j = \{\Phi^j(m) : m \in \mathbb{N}\}$ as the cumulative routing process associated with activity $j \in \mathcal{J}$. Since service times are exponential, the number of service completions of

[4] In the rehabilitation setting, the outcome is determined by the clinical reassessment that follows each treatment episode.

activity j up to time t is

$$S_j(t) \;\coloneqq\; M_j\Big(\int_0^t n_j(s)\mu_j(s)ds\Big), \quad t \geq 0, \tag{3.1}$$

where M_j is a unit rate Poisson process for $j \in \mathcal{J}$. We define the flow process $F^j = \{F^j(t) : t \geq 0\}$ associated with activity $j \in \mathcal{J}$ as follows:

$$F^j(t) \;\coloneqq\; \Phi^j\big(S_j(t)\big), \quad t \geq 0. \tag{3.2}$$

Here, $F_l^j(t)$ denotes the cumulative decrease in the number of class $l \in \mathcal{I}$ customers up to time t as a result of activity $j \in \mathcal{J}$ service completion, with negative values indicating net increases.

Each customer has a deadline, which specifies the maximum amount of time they may remain in the system. Once a customer's age in the system (i.e., sojourn or throughput time) exceeds their deadline, they become ineligible for additional service and are rejected by the system manager. As noted earlier, the system manager may also reject customers before their deadline is reached. We let δ_i denote the deadline for class $i \in \mathcal{I}$ customers. A customer's deadline is set at arrival and does not change with class transitions, so, $\delta_i = \delta_l$ for $i \in \mathcal{I}$ and $l \in \mathcal{B}_i$. In the rehabilitation setting, deadlines represent the time patients have until they age out of eligibility for the program. We let $a_i(t)$ denote the age (i.e., time elapsed since arrival) of the head-of-line customer in class $i \in \mathcal{I}$. Letting $a(t) = (a_i(t))$ for $t \geq 0$, we refer to $a = \{a(t) : t \geq 0\}$ as the age process.

The system state is described by the I-dimensional queue length process $Q = \{Q(t) : t \geq 0\}$, where $Q_i(t)$ denotes the number of class $i \in \mathcal{I}$ customers at time t. Customers may abandon when faced with excessive waiting times. To be specific, class $i \in \mathcal{I}$ customers abandon after an exponentially distributed amount of time with rate $\ell_i \geq 0$. Thus, the cumulative number of class $i \in \mathcal{I}$ customers abandoned up to time t is

$$\Gamma_i(t) \;\coloneqq\; N_i\Big(\int_0^t \ell_i Q_i(s)\, ds\Big), \quad t \geq 0,$$

where N_i is a unit-rate Poisson process. We assume that the processes E_i, N_i, M_j, and Φ^j are mutually independent across $i \in \mathcal{I}$ and $j \in \mathcal{J}$.

Assuming the system is initially empty, the queue length process evolves as follows: For $i \in \mathcal{I}$,

$$Q_i(t) \;=\; E_i(t) \;-\; \sum_{j \in \mathcal{J}} F_i^j(t) \;-\; \Gamma_i(t) \;-\; O_i(t), \quad t \geq 0, \tag{3.3}$$

where the four terms capture, respectively, the impact of external arrivals, service completions and subsequent stochastic transitions, abandonments, and rejections. A dynamic control policy (μ, n, O)

is said to be *admissible* if it is nonanticipating, i.e., decisions depend only on currently available information, and satisfies

$$O \text{ is nondecreasing with } O(0) = 0, \tag{3.4}$$

$$\mu(t) \in \mathcal{M},\ n(t) \in \mathbb{Z}_+^J, \text{ and } Q(t) \in \mathbb{Z}_+^I \text{ for } t \geq 0, \tag{3.5}$$

$$\sum_{j \in \mathcal{J}} A_{kj} n_j(t) \leq 1 \text{ for } k \in \mathcal{K} \text{ and } t \geq 0, \tag{3.6}$$

$$\sum_{j \in \mathcal{J}_i} n_j(t) \leq Q_i(t) \text{ for } i \in \mathcal{I} \text{ and } t \geq 0, \tag{3.7}$$

$$|\mu_j(t) - \mu_j(s)| \leq \Delta_j \text{ for } j \in \mathcal{J} \text{ and } s, t \geq 0, \tag{3.8}$$

$$a_i(t) \leq \delta_i \text{ for } i \in \mathcal{I} \text{ and } t \geq 0. \tag{3.9}$$

Equation (3.4) ensures that rejections are irreversible. Equations (3.5)–(3.8) ensure that the service rate and scheduling policies are well-defined and achievable. Equation (3.9) ensures that the age of the head-of-line customer in each class does not exceed its deadline.

3.3.3 Stochastic Control Problem.

A class $i(j)$ customer who waits $w_{i(j)}$ units of time in buffer $i(j)$ and is served under activity $j \in \mathcal{J}$ at service rate μ_j incurs a cost with two components: The first component, the delay cost $\hat{v}_j(\mu_j, w_{i(j)})$, represents the aggregate waiting and service cost; it is nondecreasing in waiting time and nonincreasing in service rate. The second component, the transition cost $\tilde{v}_j(\mu_j, w_{i(j)})$, represents the immediate cost of failing to achieve the desired outcome upon completion of the current episode. This cost is most consequential when the customer has reached the maximum number of episodes and must exit without achieving the desired outcome; see Chapter 8. For convenience, we aggregate these costs into a single cost function. Specifically, for each activity $j \in \mathcal{J}$, we define the (aggregate) service cost function as $v_j(\mu_j, w_{i(j)}) := \hat{v}_j(\mu_j, w_{i(j)}) + \tilde{v}_j(\mu_j, w_{i(j)})$ for $\mu_j \in \mathcal{M}_j$ and $w_{i(j)} \in \mathbb{R}_+$.

Additional costs arise from rejection and abandonment. A rejection cost v_i is incurred when a class $i \in \mathcal{I}$ customer is rejected, either due to excessive congestion or the customer's age reaching its deadline. An abandonment cost π_i is incurred when a class $i \in \mathcal{I}$ customer abandons (voluntarily) due to limited patience.

Given an admissible control policy (μ, n, O), the cumulative cost up to time t is defined as

follows:

$$V(t) \coloneqq \sum_{j\in\mathcal{J}} \int_0^t n_j(s)\mu_j(s)\upsilon_j\big(\mu_j(s), w_{i(j)}(s)\big)ds + \sum_{i\in\mathcal{I}} \pi_i\Gamma_i(t) + \sum_{i\in\mathcal{I}} \nu_i O_i(t), \quad t \geq 0. \quad (3.10)$$

The first term is a surrogate for the expected service cost; see e.g., Rubino and Ata (2009). The remaining terms are the abandonment and rejection costs. We refer to $V = \{V(t) : t \geq 0\}$ as the cumulative cost process.

Adopting the long-run average expected cost criterion, the system manager seeks to find a dynamic control policy (μ, n, O) so as to

$$\text{minimize } \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}[V(t)] \quad \text{subject to (3.4)–(3.9).} \quad (3.11)$$

Unfortunately, this formulation is a nonlinear multi-dimensional stochastic control problem. Therefore, it is not analytically tractable and suffers from the curse of dimensionality unless the number of customer classes and servers is small. Thus, we next consider a sequence of systems in the heavy traffic regime and formulate an approximating Brownian control problem, which is analytically tractable.

Chapter 4

Approximating Brownian Control Problem

This section introduces a Brownian approximation for the original control problem discussed in Chapter 3. The approximation emerges as the limit of a sequence of scaled systems under conditions of heavy traffic. The key assumption underlying the approximation is that the arrival and service rates are sufficiently large. Ultimately, the solution of the Brownian control problem is interpreted in the context of the original control problem to guide the design of our proposed policy.

4.1 Heavy-Traffic Assumption and Asymptotic Regime

We begin by defining the mapping $R : \mathcal{M} \times \mathbb{R}_+^I \to \mathbb{R}^{I \times J}$ as follows: For $\mu \in \mathcal{M}$, $w \in \mathbb{R}_+^I$, $i \in \mathcal{I}$, and $j \in \mathcal{J}$,

$$R_{ij}(\mu, w) \coloneqq \mu_j \big(\mathbf{1}_{\{j \in \mathcal{J}_i\}} - P_{ji}(\mu_j, w_{i(j)}) \big). \tag{4.1}$$

We interpret $R_{ij}(\mu, w)$ as the rate at which a single instance activity j reduces the number of class i customers, given service rate vector μ and (virtual) waiting time vector w. We next consider the following static planning problem (SPP), which ignores the randomness and all congestion-related costs and dynamics.

Static Planning Problem. The static planning problem seeks to minimize the aggregate service and rejection cost rate subject to two constraints. The first constraint ensures flow balance, i.e., input and output rates of each buffer must coincide. The second constraint imposes the capacity constraint for the servers. Specifically, it seeks a service rate vector $\mu \in \mathcal{M}$, scheduling vector $n \in \mathbb{R}_+^J$, and

rejection vector $o \in \mathbb{R}_+^I$ to

$$\text{minimize} \ \sum_{j\in\mathcal{J}} n_j \mu_j \upsilon_j(\mu_j, 0) + \sum_{i\in\mathcal{I}} \nu_i o_i \ \text{ subject to } \ R(\mu,0)\, n + o = \lambda \ \text{ and } \ An \le e, \tag{4.2}$$

where e denotes the vector of ones. If the system manager were to ignore randomness and thus all congestion-related concerns, they would use a control policy that solves the SPP (4.2). However, due to randomness, they may benefit from dynamic adjustments to this policy, which are captured by the Brownian approximation.

To develop the Brownian approximation, we consider a sequence of related systems indexed by a system parameter $r \in \mathbb{N}$. We attach a superscript r to various quantities of interest corresponding to the r-th system in this sequence. We focus on the asymptotic regime where the input rate, transition kernels, service cost functions, and rejection costs vary with r as follows: For all $r \in \mathbb{N}$, $i, l \in \mathcal{I}$, $j \in \mathcal{J}$, $\mu_j \in \mathcal{M}_j$, and $w_{i(j)} \in \mathbb{R}_+$,

$$\lambda_i^r := r\lambda_i - \sqrt{r}\zeta_i, \quad P_{jl}^r(r\mu_j, w_{i(j)}) := P_{jl}(\mu_j, w_{i(j)}), \quad \upsilon_j^r(r\mu_j, w_{i(j)}) := \upsilon_j(\mu_j, w_{i(j)}), \ \text{ and } \ \nu_i^r := \nu_i, \tag{4.3}$$

where $\lambda_i, \zeta_i, \nu_i \in \mathbb{R}$ are given constants. It follows directly from (4.3) that for $r \in \mathbb{N}$, $i \in \mathcal{I}$, and $j \in \mathcal{J}$,

$$R_{ij}^r(r\mu, w) \ = \ rR_{ij}(\mu, w).$$

We make the following assumption that is required to derive our approximation; see e.g., Harrison (2003).

Assumption 4.1 (Heavy-Traffic Assumption)**.** *The SPP (4.2) has an optimal solution* $(\mu^\star, n^\star, o^\star)$*, unique up to the service rate* μ_j *of the activities for which* $n_j^\star = 0$*. Moreover,* $An^\star = e$ *and* $\mu_j^\star \in \text{int}(\mathcal{M}_j)$ *for each activity* $j \in \mathcal{J}$ *for which* $n_j^\star > 0$*.*

We refer to $\mu^\star$, $n^\star$, and $o^\star$ as the system manager's nominal service rate, scheduling, and rejection policies, respectively. Assumption 4.1 states that the nominal scheduling policy $n^\star$ fully utilizes all server capacities, thereby placing the system in heavy traffic. We interpret $n_j^\star$ as the expected number of concurrent instances of activity j that should be undertaken, in the absence of randomness, to meet the demand. Activity $j \in \mathcal{J}$ is called basic if $n_j^\star > 0$. Let J_b denote the number of basic activities. Without loss of generality, we assume the basic activities are numbered $1, \ldots, J_b$. The remaining activities are referred to as nonbasic activities. To aid the analysis to follow, we partition the capacity consumption matrix A such that $A = [A_B \ A_N]$, where A_B has J_b

columns, and define the $(K + J - J_b) \times J$ matrix H as follows:

$$H \coloneqq \begin{bmatrix} A_B & A_N \\ 0 & -I \end{bmatrix}. \tag{4.4}$$

Under Assumption 4.1, the first J_b components of the nominal service rate vector $\mu^\star$ are uniquely determined by the SPP (4.2). Thus, we partition it as $\mu^\star = (\mu^\star_B; \mu^\star_N)$, where the J_b-dimensional subvector $\mu^\star_B$ corresponds to the basic activities and is uniquely determined, while the $(J - J_b)$-dimensional subvector $\mu^\star_N \in \mathcal{M}_N \coloneqq \{\mu_N : (\mu^\star_B; \mu_N) \in \mathcal{M}\}$ corresponds to the nonbasic activities. We treat $\mu^\star_N$ as a static design parameter and determine its value in Section 6.4. To ensure tractability, we impose the following regularity condition.

Assumption 4.2. *The service cost function υ_j and the transition kernel P_{jl} are continuously differentiable in a neighborhood of $(\mu^\star_j, 0)$ for $j \in \mathcal{J}$ and $l \in \mathcal{I}$.*

We further assume that, in the absence of randomness, it is less costly to serve customers than to reject them.

Assumption 4.3. *The optimal solution $(\mu^\star, n^\star, o^\star)$ to the SPP (4.2) satisfies $o^\star = 0$.*

It follows from Assumption 4.1 and (4.3) that if the system manager were to ignore randomness, they would choose the instantaneous service rate $r\mu^\star$. However, in the stochastic system, they may benefit from (dynamic) adjustments to the instantaneous service rate. In large balanced-flow systems in heavy traffic, the queue lengths are of order $\sqrt{r}$, which suggests that fluctuations in the service rate should also be of order $\sqrt{r}$ to effectively manage congestion. (Adjustments of smaller order are asymptotically negligible, while adjustments of a larger order would override the queue length dynamics in the limit.) Thus, we focus on instantaneous service rate vectors of the following form: For $r \in \mathbb{N}$ and some $\eta : [0, \infty) \to \mathbb{R}^J_+$,

$$\mu^r(t) \;=\; r\mu^\star \;+\; \sqrt{r}\eta(t), \quad t \geq 0; \tag{4.5}$$

see Ata (2006) and Ata and Barjesteh (2022) for similar treatments.

Moreover, as thoroughly argued in Harrison (1988, 2000, 2003), any dynamic scheduling policy worthy of consideration should satisfy the following when r is large:

$$\int_0^t n^r(s)ds \;\approx\; n^\star t, \quad t \geq 0. \tag{4.6}$$

That is, $n^\star_j t$ should give a first-order approximation to the cumulative effort allocated to activity $j \in \mathcal{J}$ up to time t. However, due to randomness, it may be beneficial to make order $1/\sqrt{r}$

deviations from these nominal rates. Thus, we define the centered and scaled allocation process $Y^r = \{Y^r(t) : t \geq 0\}$ as follows: For $j \in \mathcal{J}$,

$$Y_j^r(t) \coloneqq \sqrt{r} \int_0^t (n_j^\star - n_j^r(s)) ds, \quad t \geq 0. \tag{4.7}$$

Moreover, we define the process $\mathfrak{U}^r = \{\mathfrak{U}^r(t) : t \geq 0\}$ as follows:

$$\mathfrak{U}^r(t) \coloneqq HY^r(t), \quad t \geq 0. \tag{4.8}$$

The first K components of $\mathfrak{U}^r$ correspond to the scaled cumulative idleness of the servers, and the remaining $J - J_b$ components correspond to the scaled cumulative effort allocated to nonbasic activities up to time t.

Since queue lengths in the heavy traffic regime are of order $\sqrt{r}$, rejections should also be of the same order to effectively regulate congestion. As such, we define the scaled queue length process $Z^r = \{Z^r(t) : t \geq 0\}$ and the scaled rejection process $G^r = \{G^r(t) : t \geq 0\}$ as follows:

$$Z^r(t) \coloneqq \frac{Q^r(t)}{\sqrt{r}} \quad \text{and} \quad G^r(t) \coloneqq \frac{O^r(t)}{\sqrt{r}}, \quad t \geq 0. \tag{4.9}$$

Moreover, since queue lengths in the heavy traffic regime are of order $\sqrt{r}$ and service rates are of order r, the ages are of order $1/\sqrt{r}$. Therefore, we define the scaled age process $\alpha^r = \{\alpha^r(t) : t \geq 0\}$ as follows:

$$\alpha^r(t) \coloneqq \sqrt{r} a^r(t), \quad t \geq 0. \tag{4.10}$$

To match the scaling of age and service rates, we scale deadlines and maximum service rate variations as

$$\delta_i^r \coloneqq \frac{\delta_i}{\sqrt{r}} \quad \text{and} \quad \Delta_j^r \coloneqq \sqrt{r}\Delta_j, \tag{4.11}$$

for $i \in \mathcal{I}$ and $j \in \mathcal{J}$, where δ_i and Δ_j are given constants. We assume the abandonment rate l_i^r and the abandonment cost π_i^r do not scale with r, i.e., for $i \in \mathcal{I}$,

$$\ell_i^r \coloneqq \ell_i, \quad \text{and} \quad \pi_i^r \coloneqq \pi_i, \tag{4.12}$$

where $\ell_i, \pi_i \in \mathbb{R}_+$ are given constants. We define the centered and scaled cumulative cost process as the deviation of the cumulative cost process from that in the corresponding deterministic system.

That is,

$$\xi^r(t) \coloneqq \frac{1}{\sqrt{r}}\left(V^r(t) - r\sum_{j\in\mathcal{J}} n_j^\star \mu_j^\star \upsilon_j(\mu_j^\star, 0)\, t\right), \quad t \geq 0. \tag{4.13}$$

For ease of exposition, we next introduce several matrices and vectors that summarize the local first-order effects of service rate adjustments and congestion around the nominal operating point $(\mu^\star, 0)$, corresponding to the nominal service rate vector and zero waiting times. First, we define the $I\times J$ matrix $C = [C_{ij}]$ and the $I\times I$ matrix $D = [D_{il}]$ as follows: For $i, l \in \mathcal{I}$ and $j \in \mathcal{J}$,

$$C_{ij} \coloneqq \left(\mathbf{1}_{\{i(j)=i\}} - P_{ji}(\mu_j^\star, 0)\right) n_j^\star - n_j^\star \mu_j^\star \frac{\partial P_{ji}}{\partial \mu_j}(\mu_j^\star, 0), \quad D_{il} \coloneqq \ell_i \mathbf{1}_{\{i=l\}} - \sum_{j\in\mathcal{J}_l} \frac{n_j^\star \mu_j^\star}{\sum_{j'\in\mathcal{J}_l} n_{j'}^\star \mu_{j'}^\star} \frac{\partial P_{ji}}{\partial w_{i(j)}}(\mu_j^\star, 0), \tag{4.14}$$

where $\partial P_{ji}(\mu_j^\star, 0)/\partial \mu_j$ and $\partial P_{ji}(\mu_j^\star, 0)/\partial w_{i(j)}$ denote the partial derivatives of P_{ji} with respect to μ_j and $w_{i(j)}$, respectively, evaluated at $(\mu_j^\star, 0)$. The matrices C and D capture the effects of service rate adjustments and of abandonments and waiting times, respectively, on the drift rate of the scaled queue length process. Similarly, we define the vectors $b = (b_j)$ and $\mathfrak{g} = (\mathfrak{g}_i)$ as follows: For $i \in \mathcal{I}$ and $j \in \mathcal{J}$,

$$b_j \coloneqq n_j^\star \upsilon_j(\mu_j^\star, 0) + n_j^\star \mu_j^\star \frac{\partial \upsilon_j}{\partial \mu_j}(\mu_j^\star, 0) \quad \text{and} \quad \mathfrak{g}_i \coloneqq \pi_i \ell_i + \sum_{j\in\mathcal{J}_i} \frac{n_j^\star \mu_j^\star}{\sum_{j'\in\mathcal{J}_i} n_{j'}^\star \mu_{j'}^\star} \frac{\partial \upsilon_j}{\partial w_{i(j)}}(\mu_j^\star, 0). \tag{4.15}$$

The vector b represents the marginal cost of service rate adjustments, and $\mathfrak{g}$ represents the marginal state cost associated with abandonments and waiting times. Finally, we define the nonnegative $I\times I$ matrix Λ as

$$\Lambda \coloneqq (I - \Pi)^{-1} \operatorname{diag}\left(\frac{1}{\sum_{j\in\mathcal{J}_1} n_j^\star \mu_j^\star}, \frac{1}{\sum_{j\in\mathcal{J}_2} n_j^\star \mu_j^\star}, \ldots, \frac{1}{\sum_{j\in\mathcal{J}_I} n_j^\star \mu_j^\star}\right), \tag{4.16}$$

where $\Pi = [\Pi_{il}]$ with $\Pi_{il} = \mathbf{1}_{\{i\in\mathcal{B}_l\}}$ for $i, l \in \mathcal{I}$; Λ maps the queue length process to the age process.

4.2 Brownian Control Problem

Appendix B formally derives the Brownian approximation as r gets large. In this approximation, the centered and scaled cumulative cost process ξ^r, scaled age process α^r, scaled queue length process Z^r, centered and scaled allocation process Y^r, scaled rejection process G^r, and $\mathfrak{U}^r$, which captures

the scaled cumulative idleness and cumulative effort allocated to nonbasic activities, are replaced with their formal limits ξ, α, Z, Y, G, and $\mathfrak{U}$, which jointly satisfy the following for $t \geq 0$, and some $\mu_N^\star \in \mathcal{M}_N$ and $\Delta_0 \in \mathbb{R}^J$:

$$\xi(t) = \int_0^t b'\eta(s)ds + \int_0^t \mathfrak{g}'Z(s)ds - \sum_{j\in\mathcal{J}} \mu_j^\star \upsilon_j(\mu_j^\star, 0)Y_j(t) + \sum_{i\in\mathcal{I}} \nu_i G_i(t), \tag{4.17}$$

$$Z(t) = X(t) - \int_0^t C\eta(s)ds - \int_0^t DZ(s)ds + R(\mu^\star, 0)Y(t) - G(t), \tag{4.18}$$

$$\alpha(t) = \Lambda Z(t), \tag{4.19}$$

$$\mathfrak{U}(t) = HY(t), \tag{4.20}$$

$$\mathfrak{U} \text{ and } G \text{ are nondecreasing with } \mathfrak{U}(0) = G(0) = 0, \tag{4.21}$$

$$\Delta_0 - \Delta/2 \leq \eta(t) \leq \Delta_0 + \Delta/2, \tag{4.22}$$

$$\alpha(t) \leq \delta \text{ and } Z(t) \geq 0, \tag{4.23}$$

where X is an I-dimensional Brownian motion with drift vector $-\zeta$ and covariance matrix Σ given by

$$\Sigma := \operatorname{diag}(\lambda_1, \ldots, \lambda_I) + \sum_{j\in\mathcal{J}} \left(\mu_j^\star n_j^\star \Omega^j + \frac{n_j^\star}{\mu_j^\star} R^j(\mu^\star, 0) R^j(\mu^\star, 0)'\right),$$

where $R^j(\mu^\star, 0)$ denotes the j-th column of $R(\mu^\star, 0)$ and the $I \times I$ matrix $\Omega^j = [\Omega_{il}^j]$ is given by

$$\Omega_{il}^j := P_{ji}(\mu_j^\star, 0)\left(\mathbf{1}_{\{i=l\}} - P_{jl}(\mu_j^\star, 0)\right),$$

for $i, l \in \mathcal{I}$ and $j \in \mathcal{J}$. Equation (4.17)–(4.23) parallel the dynamics of the original control problem in Chapter 3. Equation (4.17) decomposes the cumulative cost into the cost associated with deviations from the nominal service rate policy, abandonment and waiting, deviations from the nominal scheduling policy, and rejections. Equation (4.18) describes the dynamics of the scaled queue length in terms of abandonments, waiting times, and deviations from the nominal service rate and scheduling policies. Equation (4.19) expresses the scaled age process as a linear transformation of the scaled queue length process. Equations (4.20)–(4.21) capture inherent limitations on the scheduling and rejection processes. Equation (4.22) restricts the variations in the service rate process. Equation (4.23) captures the bounds on the scaled age and queue length processes.

The vectors $\mu_N^\star$, specifying the nominal service rate of the nonbasic activities, and Δ_0, specifying the baseline service rate vector,[1] are static parameters. In contrast, the processes η, Y, and G, representing dynamic adjustments to service rate, scheduling, and rejection policies, are dynamic

[1]The baseline service rate vector in the r-th system is given by $r\mu^\star + \sqrt{r}\Delta_0$.

controls. This separation suggests a natural two-stage approach. In the first stage, the system manager chooses the static parameters $(\mu_N^\star, \Delta_0)$. In the second stage, given $(\mu_N^\star, \Delta_0)$, the system manager chooses the dynamic controls (η, Y, G). Thus, we assume $(\mu_N^\star, \Delta_0)$ is fixed and focus on the second-stage problem, referred to as the Brownian control problem. The first-stage problem is addressed in Section 6.4.

In the approximating Brownian control problem, the control policy takes the form of a tuple (η, Y, G). A control (η, Y, G) is said to be admissible if it is nonanticipating and satisfies (4.17)–(4.23). The Brownian Control Problem (BCP) can now be stated as follows: Choose a nonanticipating control (η, Y, G) so as to

$$\text{minimize } \limsup_{t\to\infty} \frac{1}{t}\,\mathbb{E}\big[\xi(t)\big] \text{ subject to (4.17)–(4.23).} \tag{4.24}$$

Remark 4.4. Since $C_{ij} = 0$ and $b_j = 0$ for $i \in \mathcal{I}$ and every nonbasic activity $j = J_b + 1, \ldots, J$, it follows from (4.17)–(4.23) that neither η_j nor $(\Delta_0)_j$ associated with nonbasic activities affects the state process or cost function of the BCP (4.24). We therefore restrict attention to control policies with $\eta_j \equiv 0$ and $(\Delta_0)_j = 0$ for all nonbasic activities $j = J_b + 1, \ldots, J$. That is, only the service rates of basic activities are dynamically adjusted, while the service rates of nonbasic activities remain fixed at their nominal values.

Although the Brownian control problem introduced in Chapter 4 is simpler than the original control problem due to the continuous nature of the state dynamics, it remains a high-dimensional stochastic control problem. We next develop an equivalent one-dimensional formulation that is analytically tractable.

Chapter 5

Equivalent Workload Formulation

This section develops an equivalent one-dimensional control problem, referred to as the equivalent workload formulation. We begin with the notion of communicating servers; see also Harrison and López (1999).

Definition 5.1. Server $k \in \mathcal{K}$ communicates directly with server $k' \in \mathcal{K}$ if there exist basic activities j and j' such that $i(j) = i(j')$, $k = k(j)$, and $k' = k(j')$. Server k communicates with server k' if there exist servers $k_1, \ldots, k_l$ such that $k_1 = k$, $k_l = k'$, and k_s communicates directly with k_{s+1} for all $s = 1, \ldots, l-1$.

Communication is an equivalence relation, meaning the servers can be partitioned into disjoint subsets such that each server communicates only with servers in the same subset. Each subset is called a server pool. We make the following standard structural assumption that enables a one-dimensional workload formulation; see Rubino and Ata (2009) and Ghamami and Ward (2013) for similar treatments.

Assumption 5.2 (Complete Resource Pooling Assumption)**.** *The optimal solution to the SPP (4.2) has $J_b = I + K - 1$ basic activities and all servers communicate.*

Under Assumption 5.2, the servers have sufficiently overlapping basic activities to effectively form a single server pool in the heavy-traffic limit. Consequently, the state process Z lies on a one-dimensional subspace, and the system dynamics can be summarized by a one-dimensional workload process. This assumption is satisfied by many systems of interest, e.g., all single-server systems and multi-server systems with sufficiently overlapping service capabilities. Because our formulation builds on Harrison (2003) and incorporates an economic objective, both conditions in Assumption 5.2 are essential, with neither implying the other. This contrasts with utilization-based formulations, in which the two are equivalent (Ata and Kumar 2005).

As a preliminary, we consider the following problem: Find $m, y \in \mathbb{R}^I$ and $g, \rho \in \mathbb{R}^{K+J-J_b}$ such that

$$m'R(\mu^\star, 0) \;=\; g'H \quad \text{and} \quad -(\mu_j^\star \upsilon_j(\mu_j^\star, 0))' \;=\; \rho'H \;+\; y'R(\mu^\star, 0). \tag{5.1}$$

The next result ensures that (5.1) admits a solution and offers insights into the structure of the solution.

Proposition 5.3. *Suppose Assumption 5.2 holds. Then, (5.1) admits a solution. In particular, the pair (m, g) is unique up to a multiplicative constant, and the pair (y, ρ) is determined up to a one-dimensional space. Moreover, there exists a solution with nonnegative ρ. For this solution, $y_i + \nu_i \geq 0$ for all $i \in \mathcal{I}$.*

Motivated by Proposition 5.3, we henceforth assume m, y, g, and ρ are fixed vectors satisfying (5.1) and ρ is nonnegative. The vectors y and ρ admit a natural dual interpretation. Specifically, y is the negative of the dual variables associated with the flow balance constraints in the SPP (4.2) when service rates are fixed at $\mu^\star$. That is, y represents the marginal cost of an additional customer in each class. The first K components of ρ are the dual variables for the capacity constraints in the SPP (4.2) and represent the marginal cost of idling servers. The remaining components of ρ represent the marginal cost of engaging in nonbasic activities.

We define the one-dimensional workload process $W = \{W(t) : t \geq 0\}$ as follows:

$$W(t) \;\coloneqq\; m'Z(t), \quad t \geq 0. \tag{5.2}$$

We interpret m_i as the workload contribution of class $i \in \mathcal{I}$. For ease of exposition, we let

$$d \;\coloneqq\; D'm. \tag{5.3}$$

Then, substituting (4.18) into (5.2) and using (5.3) yields

$$W(t) \;=\; B(t) \;-\; \int_0^t \theta(s)\, ds \;-\; \int_0^t d'Z(s)\, ds \;+\; L(t) \;-\; U(t), \quad t \geq 0, \tag{5.4}$$

where for $t \geq 0$,

$$B(t) \;\coloneqq\; m'X(t), \quad \theta(t) \;\coloneqq\; m'C\eta(t), \quad L(t) \;\coloneqq\; g^{+\prime}\mathfrak{U}(t) + m^{-\prime}G(t), \quad U(t) \;\coloneqq\; g^{-\prime}\mathfrak{U}(t) \;+\; m^{+\prime}G(t), \tag{5.5}$$

where x^+ and x^- are the componentwise positive and negative parts of x. According to (5.4), the workload process evolves as a controlled diffusion process. The first term captures the stochastic

variability inherited from X. Specifically, $B = \{B(t) : t \geq 0\}$ is a Brownian motion with infinitesimal drift $\mu := -m'\zeta$ and variance $\sigma^2 := m'\Sigma m > 0$. The second term captures the drift rate induced by the dynamic service rate adjustments. We refer to $\theta = \{\theta(t) : t \geq 0\}$ as the effective drift rate process. The third term captures the drift rate generated by congestion effects, namely, abandonments and waiting time dependent transitions; see (4.14) and (5.3). The fourth and fifth terms represent, respectively, the cumulative increase and decrease in the workload process due to the combined effects of server idleness, engaging in nonbasic activities, and rejections. The processes $L = \{L(t) : t \geq 0\}$ and $U = \{U(t) : t \geq 0\}$ are referred to as the effective upward and downward regulators.

Costs in the workload formulation arise from three sources: effective drift rate, congestion effects, and pushing the workload process. We begin with the cost associated with the effective drift rate. Let

$$\beta := b + C'y. \tag{5.6}$$

The vector β represents the marginal cost rate of effective drift rate control. The term b reflects the direct cost of service rate adjustments (see (4.15)), whereas $C'y$ captures the indirect congestion effects implied by the flow balance constraints, with y arising as the associated dual variable. For a given effective drift rate x, the system manager seeks a feasible η that achieves x at minimum cost. Accordingly, we define the optimal drift rate function $\eta^\star : \mathbb{R} \to \mathbb{R}^J$ and the effective drift rate cost function $c : \mathbb{R} \to \mathbb{R}$ as follows: For $x \in \mathbb{R}$,

$$\eta^\star(x) := \arg\min \left\{\beta'\eta : m'C\eta = x,\ |\eta - \Delta_0| \leq \Delta/2,\ \eta_j = 0 \text{ for } j > J_b,\ \eta \in \mathbb{R}^J\right\} \quad \text{and} \quad c(x) := \beta'\eta^\star(x). \tag{5.7}$$

These functions specify the optimal cost and the corresponding η for a given effective drift rate x. We define the baseline effective drift rate and the maximum variation in the effective drift rate, respectively, as

$$\Theta_0 := m'C\Delta_0 \quad \text{and} \quad \Theta := |m'C|\Delta, \tag{5.8}$$

where the absolute value is applied componentwise.

We define the minimum and maximum achievable workload levels, respectively, as follows:

$$\underline{w} := \min\{m'z : \Lambda z \leq \delta,\ z \in \mathbb{R}^I_+\} \quad \text{and} \quad \bar{w} := \max\{m'z : \Lambda z \leq \delta,\ z \in \mathbb{R}^I_+\}.$$

For each workload level $w \in [\underline{w}, \bar{w}]$, the set of admissible workload distribution vectors is given by

$$\mathcal{A}(w) \coloneqq \{z \in \mathbb{R}^I_+ : m'z = w,\ \Lambda z \leq \delta\}. \tag{5.9}$$

To characterize congestion costs, we define the I-dimensional holding cost vector h as follows:

$$h \coloneqq \mathfrak{g} + D'y, \tag{5.10}$$

where h_i represents the congestion cost associated with holding one unit of work in buffer i. The term $\mathfrak{g}$ reflects the direct congestion costs associated with abandonment and waiting (see (4.15)), while $D'y$ captures the indirect congestion effects implied by the flow balance constraints, with y arising as the associated dual variable. Distributing workload w according to $z \in \mathcal{A}(w)$ yields a congestion cost rate of $h'z$.

Finally, we characterize the cost of pushing the workload process upward and downward. Such adjustments arise through instantaneous (singular) control actions, including idling a server, engaging in a nonbasic activity, and rejecting customers. Each action changes workload at a specific rate and incurs an associated cost. The system manager seeks the least costly action to achieve a unit increase or decrease in workload. Thus, the effective costs of pushing the workload process upward and downward, respectively, are given by

$$\kappa_L \coloneqq \min_{j=1,\ldots,J-J_b+K} \frac{\rho_j}{g_j^+} \wedge \min_{i\in\mathcal{I}} \frac{y_i + \nu_i}{m_i^-} \quad \text{and} \quad \kappa_U \coloneqq \min_{j=1,\ldots,J-J_b+K} \frac{\rho_j}{g_j^-} \wedge \min_{i\in\mathcal{I}} \frac{y_i + \nu_i}{m_i^+}, \tag{5.11}$$

where $\wedge$ denotes the minimum operator, and all terms with zero denominator are omitted from the minimization. It follows from Proposition 5.3 that κ_L and κ_U are nonnegative.

In the equivalent workload formulation, a control policy takes the form of a tuple (L, U, z, θ), where L, U, z, and θ capture the impact of upward and downward pushing actions (server idleness, engaging in nonbasic activities, and rejections), dynamic scheduling, and dynamic service rate adjustments. We refer to $z : [0, \infty) \times [\underline{w}, \bar{w}] \to \mathbb{R}^I$ as the workload configuration function. A control policy (L, U, z, θ) is said to be admissible to the workload formulation if it is nonanticipating

and satisfies the following for $t \geq 0$:

$$\Xi(t) \;=\; \int_0^t c(\theta(s))ds + \int_0^t h'z(s, W(s))ds \;+\; \kappa_L L(t) \;+\; \kappa_U U(t), \tag{5.12}$$

$$W(t) \;=\; B(t) \;-\; \int_0^t \theta(s)\, ds \;-\; \int_0^t d'z\big(s, W(s)\big)\, ds \;+\; L(t) \;-\; U(t), \tag{5.13}$$

$$z(t, W(t)) \;\in\; \mathcal{A}(W(t)), \tag{5.14}$$

$$L \text{ and } U \text{ are nondecreasing with } L(0) = U(0) = 0, \tag{5.15}$$

$$\Theta_0 - \Theta/2 \;\leq\; \theta(t) \;\leq\; \Theta_0 + \Theta/2, \tag{5.16}$$

where B is a Brownian motion with infinitesimal drift μ and variance σ^2. The equivalent workload formulation (EWF) can now be stated as follows: Choose an admissible control policy (L, U, z, θ) so as to

$$\text{minimize } \limsup_{t\to\infty} \frac{1}{t}\, \mathbb{E}\big[\Xi(t)\big] \text{ subject to (5.12)–(5.16).} \tag{5.17}$$

The following result establishes the equivalence of the BCP (4.24) and the EWF (5.17).

Proposition 5.4. *For every admissible policy (η, Y, G) for the BCP (4.24), there exists an admissible policy (L, U, z, θ) for the EWF (5.17), and its cost is less than or equal to that of the policy (η, Y, G) for the BCP (4.24) minus the constant $y'\zeta$. Conversely, for every admissible policy (L, U, z, θ) for the EWF (5.17), there exists an admissible policy (η, Y, G) for the BCP (4.24), and its cost is equal to that of the policy (L, U, z, θ) for the EWF (5.17) plus the constant $y'\zeta$.*

As preliminary to solving the workload formulation, we next present a useful technical result.

Lemma 5.5. *For each nonbasic activity $j = J_b + 1, \ldots, J$, we have that $\beta_j = (m'C)_j = 0$. Moreover, the vector β is a scalar multiple of $m'C$.*

Lemma 5.5 asserts that adjusting the service rates of nonbasic activities neither impacts the workload nor incurs a cost. Thus, we restrict attention to policies satisfying $\eta_j \equiv 0$ for $j = J_b + 1, \ldots, J$. Lemma 5.5 also asserts that the cost per unit of effective drift rate, $\beta_j/(m'C)_j$, is the same for all basic activities $j = 1, \ldots, J_b$.

Chapter 6

Solution to the Equivalent Workload Formulation

This section solves the workload formulation. To minimize technical complexity, we restrict attention to stationary Markov control policies, under which the workload configuration $z(t, W(t))$ and the effective drift rate $\theta(t)$ at time t depend on the history only through the current workload $W(t)$. Thus, in what follows, we write $z(W(t))$ and $\theta(W(t))$ in place of $z(t, W(t))$ and $\theta(t)$, respectively. The remainder of this section is organized as follows. Section 6.1 introduces a class of barrier policies. Section 6.2 introduces the associated Bellman equation. Section 6.3 solves the Bellman equation, and derives an optimal solution (which is a barrier policy) to the workload formulation. Finally, Section 6.4 addresses the static design parameters.

6.1 Barrier Policies

Definition 6.1. Given $l, u \in [\underline{w}, \bar{w}]$ such that $l < u$, we call an admissible control policy (L, U, z, θ) a barrier policy (with a lower barrier at l and an upper barrier at u) if it satisfies $W(t) \in [l, u]$ for $t \geq 0$, and

$$\int_0^t \mathbf{1}_{\{W(s)>l\}}\, dL(s) \;=\; 0 \quad \text{and} \quad \int_0^t \mathbf{1}_{\{W(s)<u\}}\, dU(s) \;=\; 0, \quad t \geq 0. \tag{6.1}$$

The processes L and U enforce the lower and upper reflecting barriers at l and u, respectively. We next establish a method for computing the long-run average expected cost of a barrier policy. As a preliminarily, let $C^n[l, u]$ for $n \in \mathbb{N}$ denote the space of functions $f : [l, u] \to \mathbb{R}$ that are n-times continuously differentiable. Then, given functions $z : \mathbb{R} \to \mathbb{R}^I$ and $\theta : \mathbb{R} \to \mathbb{R}^J$ and $\gamma \in \mathbb{R}$, define

the differential operator $\Gamma_{z,\theta}$ as follows:

$$\Gamma_{z,\theta} f(w) \coloneqq \frac{1}{2}\sigma^2 f''(w) + \mu f'(w) - \theta(w) f'(w) - d' z(w) f'(w), \quad (f, w) \in C^2[l, u] \times [l, u]. \tag{6.2}$$

Then, we consider the following second-order differential equation:

$$\Gamma_{z,\theta} f(w) + c(\theta(w)) + h' z(w) = \gamma, \quad w \in (l, u), \tag{6.3}$$

subject to the boundary conditions

$$f'(l) = -\kappa_L \quad \text{and} \quad f'(u) = \kappa_U. \tag{6.4}$$

Proposition 6.2. *Consider the barrier policy (L, U, z, θ) with a lower barrier at l and an upper barrier at u. Suppose there exist $\gamma \in \mathbb{R}$ and $f \in C^2([l, u])$ that jointly satisfy (6.3)–(6.4), then*

$$\lim_{t\to\infty} \frac{1}{t}\mathbb{E}\left[\int_0^t c(\theta(W(s)))ds + \int_0^t h' z(W(s))ds + \kappa_L L(t) + \kappa_U U(t) \right] = \gamma.$$

6.2 Bellman Equation

Proposition 6.2 and smooth pasting arguments (Harrison 2013, Section 7.7) lead naturally to the following Bellman equation: Find constants $l, u \in [\underline{w}, \bar{w}]$, $\gamma \in \mathbb{R}$ and a function $f \in C^2([l, u])$ that satisfy $l \le u$ and

$$\min_{\substack{x\in[\Theta_0-\Theta/2,\Theta_0+\Theta/2] \\ z\in\mathcal{A}(w)}} \left\{ \frac{1}{2}\sigma^2 f''(w) + \mu f'(w) - x f'(w) - d' z f'(w) + c(x) + h' z \right\} = \gamma, \quad w \in [l, u], \tag{6.5}$$

subject to the boundary conditions

$$f'(l) = -\kappa_L \quad \text{and} \quad f'(u) = \kappa_U, \tag{6.6}$$

and the smooth pasting conditions

$$f''(l) = 0 \text{ if } l > \underline{w} \quad \text{and} \quad f''(u) = 0 \text{ if } u < \bar{w}. \tag{6.7}$$

The function f is referred to as the relative value function. Its derivative can be interpreted as the marginal cost of an additional unit of workload. The Bellman equation characterizes a candidate

effective drift rate and workload distribution by balancing the marginal cost of workload against the costs of drift rate control and congestion. We interpret γ as a guess at the minimum long-run average cost and l and u as the lower and upper reflecting barriers, respectively, to be imposed on the workload process. The first-order boundary conditions in (6.6) reflect the cost implications of enforcing the reflecting barriers, and the second-order boundary conditions in (6.7), known as smooth pasting conditions, ensure that the barriers are chosen optimally. Note that the Bellman equation is introduced primarily to motivate our solution approach, and the properties we need from it will be proved from first principles.

Since the minimand in (6.5) is additively separable in x and z, and the feasible sets are independent, the minimization over x and z decouples. To address the minimization over x, we define $\psi : \mathbb{R} \to \mathbb{R}$ as follows:

$$\psi(y) \;:=\; \min\{c(x) - xy : x \in [\Theta_0 - \Theta/2, \Theta_0 + \Theta/2]\}, \quad y \in \mathbb{R}. \tag{6.8}$$

To address the minimization over z, we define the effective state cost function $\phi : [\underline{w}, \bar{w}] \times \mathbb{R} \to \mathbb{R}$ as follows:

$$\phi(w, y) \;:=\; \min\left\{h'z - d'zy : z \in \mathcal{A}(w)\right\}, \quad w \in [\underline{w}, \bar{w}],\ y \in \mathbb{R}. \tag{6.9}$$

The following result provides useful structural properties for the functions ϕ and ψ.

Lemma 6.3. *The following hold:*

(i) The function ψ is concave, continuous, piecewise linear, and almost everywhere differentiable.

(ii) For $y \in \mathbb{R}$, the mapping $w \mapsto \phi(w, y)$ is convex, continuous, piecewise linear, and almost everywhere differentiable on $[\underline{w}, \bar{w}]$. Moreover, the mapping $y \mapsto \phi(w, y)$ is uniformly Lipchitz in w.

The Bellman equation (6.5)–(6.7) is equivalent to a first-order differential equation by letting $v = f'$. Thus, it can be rewritten as follows: Find $l, u \in [\underline{w}, \bar{w}]$, $\gamma \in \mathbb{R}$ and $v \in C^1([l, u])$ that satisfy $l \leq u$ and

$$\frac{1}{2}\sigma^2 v'(w) \;+\; \mu v(w) \;+\; \psi(v(w)) \;+\; \phi(w, v(w)) \;=\; \gamma, \quad w \in [l, u], \tag{6.10}$$

subject to the boundary conditions

$$v(l) = -\kappa_L \quad \text{and} \quad v(u) = \kappa_U, \tag{6.11}$$

and the smooth pasting conditions

$$v'(l) = 0 \text{ if } l > \underline{w} \quad \text{and} \quad v'(u) = 0 \text{ if } u < \bar{w}. \tag{6.12}$$

6.3 Solution to the Bellman Equation and the Equivalent Workload Formulation

This section derives a solution to the Bellman equation, and leverages it to construct a candidate barrier policy. Then, it proves that this policy is optimal for the workload formulation. We begin by identifying a lower bound for the long-run average cost γ of any solution to the Bellman equation (6.10)–(6.12). Let

$$\underline{\gamma} \coloneqq \min\left\{ -\mu\kappa_L + \psi(-\kappa_L) + \phi(w, -\kappa_L) : w \in [\underline{w}, \bar{w}] \right\}. \tag{6.13}$$

The next result shows that any solution to the Bellman equation satisfies $\gamma \in [\underline{\gamma}, \infty)$ and establishes several properties of the effective state cost function ϕ and the constant l of any solution to the Bellman equation.

Lemma 6.4. *Suppose (l, u, γ, v) solves the Bellman equation (6.10)–(6.12). Then, we have that $\gamma \in [\underline{\gamma}, \infty)$. Moreover, if $l > \underline{w}$, the mapping $w \mapsto \phi(w, -\kappa_L)$ is nonincreasing on $[\underline{w}, l]$, and*

$$-\mu\kappa_L + \psi(-\kappa_L) + \phi(l, -\kappa_L) = \gamma. \tag{6.14}$$

Motivated by this result, we henceforth restrict attention to $\gamma \in [\underline{\gamma}, \infty)$, and for each $\gamma \in [\underline{\gamma}, \infty)$, we define

$$\mathcal{L}_\gamma \coloneqq \left\{ w \in [\underline{w}, \bar{w}] : -\mu\kappa_L + \psi(-\kappa_L) + \phi(w, -\kappa_L) = \gamma,\ w \mapsto \phi(w, -\kappa_L) \text{ is nonincreasing on } [\underline{w}, l] \right\}.$$

It follows from the convexity of $w \mapsto \phi(w, -\kappa_L)$ (by Lemma 6.3) that $\mathcal{L}_\gamma$ is either an empty set, a singleton, or a closed interval. Then, for $\gamma \in [\underline{\gamma}, \infty)$, we define

$$l_\gamma \coloneqq \begin{cases} \min \mathcal{L}_\gamma, & \text{if } \mathcal{L}_\gamma \neq \varnothing, \\ \underline{w}, & \text{otherwise,} \end{cases} \tag{6.15}$$

and interpret it as a lower bound on l for any solution to the Bellman equation with long-run average cost γ.

Lemma 6.5. *The mapping $\gamma \mapsto l_\gamma$ is continuous and nonincreasing on $[\underline{\gamma}, \infty)$.*

For each $\gamma \in [\underline{\gamma}, \infty)$, consider the following initial value problem (IVP):

$$\frac{1}{2}\sigma^2 v'(w) + \mu v(w) + \psi(v(w)) + \phi(w, v(w)) = \gamma, \quad w \in [l_\gamma, \bar{w}], \tag{6.16}$$

subject to the boundary condition

$$v(l_\gamma) = -\kappa_L. \tag{6.17}$$

Appendix C.1 establishes the existence and uniqueness of the solution $v_\gamma \in C^1([l_\gamma, \bar{w}])$ to the IVP (6.16)–(6.17). It also establishes several useful properties of the solution that are used in the subsequent analysis. For ease of exposition, we next define the function $V : [\underline{\gamma}, \infty) \to \mathbb{R}$ as follows:

$$V(\gamma) \coloneqq \max\left\{v_\gamma(w) : w \in [l_\gamma, \bar{w}]\right\}, \quad \gamma \in [\underline{\gamma}, \infty). \tag{6.18}$$

For each $\gamma \in [\underline{\gamma}, \infty)$, the quantity $V(\gamma)$ denotes the maximum value attained by v_γ on $[l_\gamma, \bar{w}]$.

Lemma 6.6. *The function V is continuous and strictly increasing. Moreover, $V(\underline{\gamma}) < \kappa_U$ and $\lim_{\gamma \to \infty} V(\gamma) = \infty$.*

By Lemma 6.6 that there exists a unique $\gamma^\star \in (\underline{\gamma}, \infty)$ such that $V(\gamma^\star) = \kappa_U$; the parameter $\gamma^\star$ can be readily computed using a bisection search. Given $\gamma^\star$, we define $l^\star, u^\star \in \mathbb{R}$ and $v^\star : [l^\star, u^\star] \to \mathbb{R}$ as follows:

$$l^\star \coloneqq l_{\gamma^\star}, \quad u^\star \coloneqq \min\left\{w : v_{\gamma^\star}(w) = \kappa_U,\ w \in [l_{\gamma^\star}, \bar{w}]\right\}, \quad \text{and} \quad v^\star(w) \coloneqq v_{\gamma^\star}(w), \quad w \in [l^\star, u^\star]. \tag{6.19}$$

The next result establishes existence and some key properties of a solution to the Bellman equation (6.10)–(6.12).

Theorem 6.7. *The tuple $(l^\star, u^\star, \gamma^\star, v^\star)$ solves the Bellman equation (6.10)–(6.12). Moreover, $-\kappa_L \leq v^\star(w) \leq \kappa_U$ for $w \in [l^\star, u^\star]$. Furthermore, if $u^\star < \bar{w}$, then the mapping $w \mapsto \phi(w, \kappa_U)$ is nondecreasing on $[u^\star, \bar{w}]$.*

To recover a solution to the original Bellman equation (6.5)–(6.7), we let

$$f^\star(w) \coloneqq \int_{l^\star}^{w} v^\star(x)dx, \quad w \in [l^\star, u^\star]. \tag{6.20}$$

Corollary 6.8. *The tuple $(l^\star, u^\star, \gamma^\star, f^\star)$ given by (6.19)–(6.20) solves the Bellman equation (6.5)–(6.7).*

We next propose a candidate barrier policy for the workload formulation. The candidate policy has lower and upper barriers at $l^\star$ and $u^\star$, with effective drift rate and workload configuration functions given by

$$\theta^\star(w) \coloneqq \arg\min\{c(x) - xv^\star(w) : x \in [\Theta_0 - \Theta/2, \Theta_0 + \Theta/2]\}, \quad w \in [l^\star, u^\star], \tag{6.21}$$

$$z^\star(w) \coloneqq \arg\min\{h'z - d'zv^\star(w) : z \in \mathcal{A}(w)\}, \quad w \in [l^\star, u^\star]. \tag{6.22}$$

The next result characterizes the candidate effective drift rate function $\theta^\star$ in closed-form. By Lemma 5.5, the ratios $\beta_j/(m'C)_j$ are identical for all basic activities $j = 1, \ldots, J_b$; we denote this common value by θ_0.

Lemma 6.9. *We have that*

$$\theta^\star(w) = \begin{cases} \Theta_0 - \Theta/2, & \text{if } v^\star(w) < \theta_0, \\ \Theta_0 + \Theta/2, & \text{otherwise.} \end{cases} \tag{6.23}$$

We illustrate the structure of the workload configuration function $z^\star$ through an example in Section 8.2. The following result provides an optimal solution to the workload formulation.

Theorem 6.10. *The barrier policy $(L^\star, U^\star, z^\star, \theta^\star)$ with lower barrier at $l^\star$, upper barrier at $u^\star$, effective drift rate function $\theta^\star$ given by (6.23), and workload configuration function $z^\star$ given by (6.22) is an optimal solution to the EWF (5.17), and it has long-run average cost of $\gamma^\star$.*

Corollary 6.11. *There exists an optimal policy for the BCP (4.24) with long-run average cost $\gamma^\star + y'\zeta$.*

6.4 First-Stage Problem

As discussed in Chapter 4, the Brownian approximation decomposes into a two-stage formulation. In the first stage, we select the static parameters $(\mu_N^\star, \Delta_0)$, which specify the nominal service rate of the nonbasic activities and the baseline service rate vector. In the second stage, conditional on $(\mu_N^\star, \Delta_0)$, we choose the dynamic control policy (η, Y, G). Given $(\mu_N^\star, \Delta_0)$, Corollary 6.11 identifies an optimal dynamic policy $(\eta^\star, Y^\star, G^\star)$ for the BCP (4.24). The remaining question is how to select the static parameters.

The BCP (4.24) is equivalent to the EWF (5.17) by Proposition 5.4. Moreover, the baseline service rate vector $\Delta_0 \in \mathbb{R}^J$ affects the EWF (5.17) only through the baseline effective drift rate $\Theta_0 \in \mathbb{R}$; see (5.8). Furthermore, Remark 4.4 shows that the baseline service rates of the nonbasic

activities may be set to zero. Therefore, in the first-stage problem, we may restrict attention to baseline service rate vectors of the form

$$(\Delta_0)_j = \begin{cases} \Theta_0/m'Ce, & j = 1, \ldots, J_b, \\ 0, & \text{otherwise} \end{cases} \tag{6.24}$$

for $j \in \mathcal{J}$, i.e., vectors with a common baseline service rate for all basic activities and zero baseline rate for all nonbasic activities. Hence, searching over $(\mu_N^\star, \Delta_0)$ is equivalent to searching over $(\mu_N^\star, \Theta_0)$.

There are two natural approaches to selecting $(\mu_N^\star, \Theta_0)$. Under the first, we minimize the long-run average cost of the BCP (4.24). Specifically, letting $\gamma^\star(\mu_N^\star, \Theta_0)$ denote the long-run average cost of the EWF (5.17) and $y(\mu_N^\star, \Theta_0)$ the vector y defined in (5.1), conditional on $(\mu_N^\star, \Theta_0)$, we select $(\mu_N^\star, \Theta_0)$ that minimizes $\gamma^\star(\mu_N^\star, \Theta_0) + y(\mu_N^\star, \Theta_0)'\zeta$. Under the second approach, we choose $(\mu_N^\star, \Theta_0)$ using a simulation-based optimization of the long-run average cost of the policy proposed in Chapter 7 for the original control problem.

Chapter 7

Proposed Policy

This section proposes a control policy for the problem introduced in Chapter 3 by interpreting the solution to the workload formulation in the context of the original control problem. The proposed policy consists of three components: scheduling, service rate, and rejection. Recall that we consider a sequence of systems indexed by r, whose formal limit is the BCP (4.24). The original control problem correspond to a particular choice of r. Thus, to describe the proposed policy, we first fix r and use it to unscale the processes of interest.

Theorem 6.10 establishes that the optimal solution to the workload formulation has the following structure. First, the workload is constrained to $[l^\star, u^\star]$ using the effective upward and downward regulators L and U. Second, for each $w \in [l^\star, u^\star]$, workload is distributed across classes according to the optimal workload configuration $z^\star(w)$. In doing so, the (scaled) age process is maintained below the (scaled) deadline vector. Third, for each $w \in [l^\star, u^\star]$, the effective drift rate $\theta^\star(w)$ is applied.

The proposed policy implements this structure in the original system. Since the proposed policy depends on workload, we begin by defining the unscaled workload process $W^r(t) \coloneqq \sum_{i\in\mathcal{I}} m_i Q_i^r(t)$ for $t \geq 0$. While the workload process in the workload formulation lives in $[l^\star, u^\star]$, the scaled nominal workload $W^r(t)/\sqrt{r}$ may leave this interval. Thus, we define the projected workload as $\mathscr{W}^r(t) \coloneqq l^\star \vee (W^r(t)/\sqrt{r} \wedge u^\star)$ for $t \geq 0$, which projects $W^r(t)/\sqrt{r}$ into $[l^\star, u^\star]$, and present the policy in terms of the projected workload process.

Barrier Control. The solution to the workload formulation imposes reflecting barriers at $l^\star$ and $u^\star$ through the regulators L and U. In the original system, these regulators correspond to control actions that increase or decrease the workload when the projected workload $\mathscr{W}^r(t)$ reaches $l^\star$ and $u^\star$, respectively. To construct these actions, we identify controls attaining the minimum in (5.11) for κ_L and κ_U, which we refer to as the upward and downward controls, respectively. Each control takes one of three forms: (1) idling a server by suspending all activities undertaken by the server;

(2) engaging in a nonbasic activity, by setting $n_j(t) = 1$ when $Q_{i(j)}(t) > 0$; or (3) rejecting the head-of-line customer of a given class. The proposed policy applies the upward control when $\mathscr{W}^r(t)$ reaches $l^\star$, and applies the downward control when $\mathscr{W}^r(t)$ reaches $u^\star$.

Service Rate Policy. The service rate policy sets the service rates according to the drift rate prescribed by the optimal solution to the BCP (4.24), using (4.5), (5.7), and (6.23). Specifically, it chooses

$$\mu^r(t) = r\mu^\star + \sqrt{r}\eta^\star\big(\theta^\star(\mathscr{W}^r(t))\big) = \begin{cases} r\mu^\star + \sqrt{r}\Delta_0 - \sqrt{r}\,\mathrm{sign}(m'C) \odot \Delta/2, & \text{if } v^\star(\mathscr{W}(t)) < \theta_0, \\ r\mu^\star + \sqrt{r}\Delta_0 + \sqrt{r}\,\mathrm{sign}(m'C) \odot \Delta/2, & \text{otherwise,} \end{cases} \tag{7.1}$$

where $\odot$ denotes componentwise (Hadamard) multiplication and $\theta_0 = \beta_j/(m'C)_j$ for $j = 1, \ldots, J_b$; see Lemma 5.5. In effect, the proposed service rate policy alternates between two service rate vectors: one used when the value function is low and another when the value function is high.

Rejection Policy. The rejection policy rejects customers in two cases: (1) when a customer's age exceeds its deadline, and (2) when rejection is prescribed by the workload formulation, namely, when $\mathscr{W}(t)$ reaches $l^\star$ or $u^\star$ and the corresponding upward or downward control specifies rejecting customers of a given class.

Scheduling Policy. The proposed scheduling policy seeks to align the workload distribution with the optimal workload configuration $z^\star$ while reducing deadline violations and avoiding unintended server idleness. As is customary in the heavy traffic literature, we introduce age slacks and safety stocks, denoted by a_i and s_i for $i \in \mathcal{I}$, respectively, to account for the gap between the diffusion-scaled queue lengths prescribed by the optimal workload configuration and the dynamics of the original system; the age slacks and safety stocks can be calibrated through simulation (Plambeck 2004, Ata and Barjesteh 2022). The scheduling policy implements a state-dependent target-tracking rule, treating $\sqrt{r}z^\star(\mathscr{W}^r(t)) + s$ as a moving target for the queue length vector and steering the system toward this target, while maintaining age slacks, i.e., $\Lambda Q^r(t) + \mathrm{a} \le \delta^r$.

In the policy described below, when the policy chooses to serve a class $i \in \mathcal{I}$ customer, it selects one of the basic activities serving class i (excluding suspended activities), prioritizing those whose server would otherwise experience idleness, i.e., those for which the remaining unassigned customers it can serve under basic activities are insufficient to keep it fully utilized. Among such activities, one is chosen with probability proportional to its nominal rate $n_j^\star$. If no such activity is available, the customer is not served.

The proposed policy operates in three stages. First, it prioritizes classes in danger of violating their age deadline, i.e., $(\Lambda Q^r(t))_i + \mathrm{a}_i \ge \delta_i^r$. Among such classes, it ranks them in descending order of

rejection cost per unit of work, $(y_i + \nu_i)/|m_i|$, and serves the $\lceil (\Lambda^{-1}(\Lambda Q^r(t) + \mathrm{a} - \delta^r)^+)_i \rceil$ head-of-line customers in each class.[1] Second, it aims to align the workload distribution with the desired target by serving classes whose queue lengths exceed the desired target, i.e., $Q_i^r(t) > \sqrt{r} z_i^\star(\mathscr{W}^r(t)) + s_i$. It ranks these classes in descending order of effective state cost per unit of work $(h_i - d_i v^\star(\mathscr{W}^r(t)))/m_i$ and serves the $Q_i^r(t) - \sqrt{r} z_i^\star(\mathscr{W}^r(t)) - s_i$ head-of-line customers in each class. Third, it uses the remaining capacity to decrease the total effective state cost. It ranks all classes in descending order of effective state cost per unit of work $(h_i - d_i v^\star(\mathscr{W}^r(t)))/m_i$ and serves remaining customers in each class, subject to available server capacity.

[1] Alternatively, one may prioritize customers in danger of exceeding their age deadline according to their age (referred to as the least absolute slack rule) or their age relative to the deadline (referred to as the least relative slack rule); see Plambeck et al. (2001).

Chapter 8

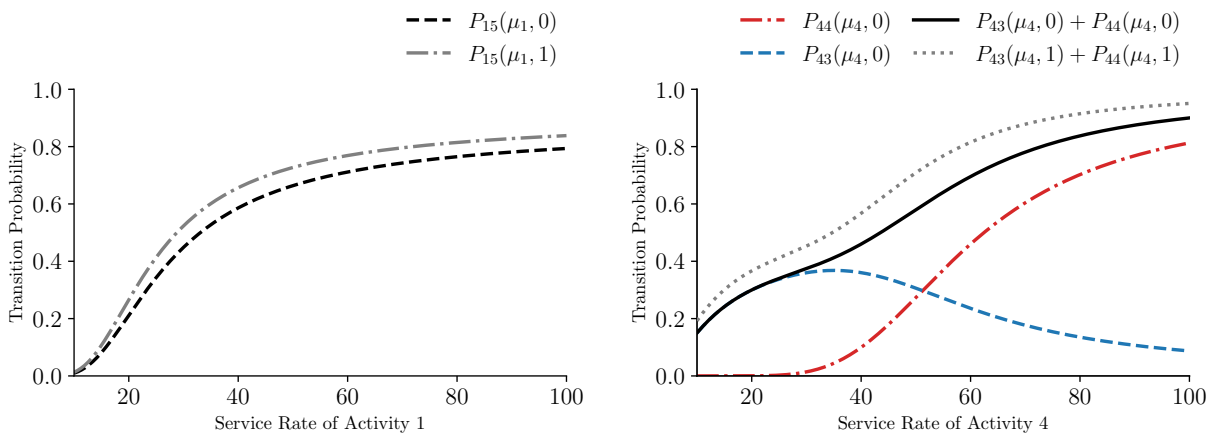
Simulation Study

This section evaluates the effectiveness of our proposed policy and illustrates several structural properties of dynamic scheduling and service rate control in multiclass multi-episode service systems with waiting time dependent transitions. We consider a service system, e.g., a rehabilitation center, with $K = 2$ servers, where external arrivals fall under two types, referred to as low and high; see Figure 3.2. Customers may receive up to two episodes of service. Upon arrival, low-type customers enter buffer 1 and, if additional service is required after the first episode, transition to buffer 5. High-type customers enter buffer 2 and if additional service is required after the first episode, transition either to buffer 3 (if they become a low-type) or buffer 4 (if they remain a high-type). Consequently, the system has $I = 5$ customer classes. Each server can serve each customer class, which yields $J = 10$ activities; see Figure 3.2 for their numbering.

The rest of this section is organized as follows. Section 8.1 specifies the system parameters in the base case. Section 8.2 examines the structure of the proposed policy, and Section 8.3 describes the benchmark policies used for comparison. Section 8.4 reports the simulation results and discusses their managerial implications. Finally, Section 8.5 extends the analysis to a setting with non-preemptive service and discrete service times, with each customer's service rate held constant over each episode.

8.1 System Parameters in the Base Case

This section specifies the base case parameters. To assess robustness, a sensitivity analysis is conducted in Section 8.4 with respect to several key model parameters. One unit of time corresponds to a year. We set the external arrival rates to $\lambda_1 = 245$, $\lambda_2 = 105$, $\lambda_3 = \lambda_4 = \lambda_5 = 0$, the set of admissible service rate vectors to $\mathcal{M} = \{\mu \in \mathbb{R}^J : 10 \leq \mu_j \leq 100, j \in \mathcal{J}\}$, and the maximum



(a) The probability that a low-type customer transitions from buffer 1 to buffer 5 after the first episode of service, for waiting times of zero and one unit of time.

(b) The probabilities that a high-type customer transitions from buffer 2 to buffers 3 or 4, or remains in the system after the first episode of service, for waiting times of zero and one unit of time.

Figure 8.1: Transition probability kernels for low-type (left panel) and high-type (right panel) customers as functions of the service rate. Higher service rates and longer waiting times increase the likelihood of requiring additional service.

service rate variation to $\Delta_j = 0.32$ for $j \in \mathcal{J}$.[1] The transition kernels take the following logistic form: For $i \in \mathcal{I}$ and $j \in \mathcal{J}$,

$$P_{jl}(\mu_j, w_{i(j)}) = \begin{cases} e^{x_{1j}}/(1 + e^{x_{1j}}), & j \in \{1, 2\},\ l = 5, \\ e^{x_{1j}}/(1 + e^{x_{1j}} + e^{x_{2j}}), & j \in \{3, 4\},\ l = 3, \\ e^{x_{2j}}/(1 + e^{x_{1j}} + e^{x_{2j}}), & j \in \{3, 4\},\ l = 4, \\ 0, & \text{otherwise}, \end{cases} \tag{8.1}$$

where $x_{kj} = a_{kj} + b_{kj}/\mu_j + c_{kj} w_{i(j)}$ for $k = 1, 2$ and $j \in \mathcal{J}$, and the parameters a_{kj}, b_{kj}, c_{kj} are given in Table D.1; see Figure 8.1 for an illustration. Server 1 is more effective at serving customers initially entering as low-type, while server 2 is more effective at serving customers initially entering as high-type. Each specialized server achieves the same transition probabilities as the other with 10% lower service rates.

The probability that a customer served by activity $j = 5, \ldots, 10$ (i.e., in a second episode of service) does not achieve the desired service outcome takes the following logistic form:

$$\hat{P}_j(\mu_j, w_{i(j)}) = e^{x_{1j}}/(1 + e^{x_{1j}}), \tag{8.2}$$

where $x_{1j} = a_{1j} + b_{1j}/\mu_j + c_{1j} w_{i(j)}$ and the parameters a_{1j}, b_{1j}, c_{1j} are given in Table D.1.

[1] This corresponds to a total (external) arrival rate of 350 customers per year, 70% of which are low-type customers. The average service time (for each episode) is between 0.5 and 5 weeks.

Table 8.1: Nominal scheduling and service rate vectors, and service costs at the nominal service rates and zero waiting time for the base case. Service rates of nonbasic activities are optimized as described in Section 6.4.

j	1	2	3	4	5	6	7	8	9	10
$n_j^\star$	5.76	4.63	0	2.92	0	2.08	0	0.37	4.24	0
$\mu_j^\star$	0.25	0.22	1	0.36	1	0.19	1	0.15	0.19	0.17
$v_j^\star(\mu_j^\star, 0)$	0.26	0.30	0.03	0.15	3.25	1.08	6.46	1.19	1.19	1.17

We set the abandonment rates to $\ell_i = 0$ and the deadlines to $\delta_i = 2$ for $i \in \mathcal{I}$. The abandonment and rejection costs are given in Table D.2.[2] The delay cost function is given by

$$\hat{v}_j(\mu_j, w_{i(j)}) = \left(\frac{10}{\mu_j}\right)^{\frac{3}{2}} e^{c_{3j} w_{i(j)}}, \tag{8.3}$$

where $c_{3j} = \ln(2)/2 \approx 0.35$ for $j \in \mathcal{J}$; see Figure D.1 (in Appendix D) for an illustration.[3] In this example, the transition costs capture the expected cost associated with failing to achieve the desired service outcome after the second (and final) episode. Accordingly, for activities $j = 1, \ldots, 4$ (the first episode), the transition cost is zero. For activities $j = 5, \ldots, 10$ (the second episode), the transition cost is $\nu_{i(j)} \hat{P}_j(\mu_j, w_{i(j)})$, where $\hat{P}_j(\mu_j, w_{i(j)})$ is the probability that a customer does not achieve the desired service outcome after the second (and final) episode; see (8.2).

8.2 Policy Parameters and the Structure of the Proposed Policy

This section presents the parameters of our proposed policy and highlights its key structural features. We set the system parameter to $r = 100$. The nominal scheduling and service rate vectors are reported in Table 8.1, and the nominal rejection vector is identically zero. As described in Section 6.4, the nominal service rate vector for nonbasic activities $\mu_N^\star$ is the minimizer of the long-run average cost of the BCP (4.24). Under the nominal policy, both servers are fully utilized and Assumptions 4.1–5.2 are satisfied. The vectors m, g, y, ρ, h, and d associated with the workload formulation are reported in Tables D.2–D.3 (in Appendix D). The minimum and maximum achievable workload levels are $\underline{w} = 0$ and $\bar{w} = 20$, respectively. The effective costs of pushing the workload process upward and downward are $\kappa_L = 0$ and $\kappa_U = 5.90$, respectively.

To begin with, we set the baseline service rate vector Δ_0 to zero. As we show later (in Table D.4), this choice is near optimal in the base case. The effective downward regulator corresponds

[2]Abandonment costs are half the rejection costs, and rejection costs are higher for high-type and new customers; see Table D.2.

[3]The delay cost is increasing and strictly convex in the expected service time, it is normalized such that its maximum at the waiting time of zero is one, and it doubles at the waiting time of two units of time.

to rejecting class 5 customers, while the effective upward regulator corresponds to engaging in nonbasic activity 10 (server 2 serving buffer 5); see Figure 3.2. The optimal solution to the workload formulation imposes a lower barrier at $l^\star = 0$ and an upper barrier at $u^\star = 7.91$; see Figure D.2 for an illustration of the associated value function.

The age slacks a_i and safety stocks s_i are set to zero for all $i \in \mathcal{I}$. Although tuning these parameters may further improve performance, the proposed policy performs well (for this example) without calibration.

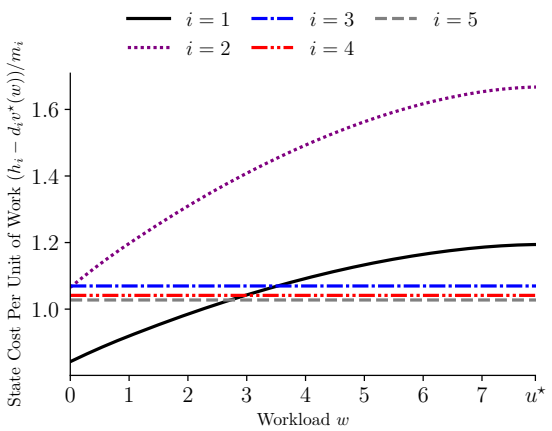
Workload Configuration and the Scheduling Policy. Holding workload in buffer $i \in \mathcal{I}$ affects the workload formulation in two distinct ways. First, it incurs an immediate holding cost at rate h_i/m_i per unit of work; see (5.12). The holding cost h_i captures the aggregate immediate cost in the original system, comprising abandonment costs, delay costs, and the cost associated with failing to achieve the desired outcome after the final episode (captured by the transition cost). In the base case, abandonment rates are zero and the transition cost is incurred only after the second (and final) episode, so the holding cost of the initial buffers (buffers 1–2) is lower than that of the secondary buffers (buffers 3–5); see Table D.2.

Second, holding workload in buffer i affects the evolution of the workload process through the drift term $-d_i z_i(w)$; see (5.13). This term captures the aggregate effect of abandonment and waiting time dependent transitions. In the base case, abandonment rates are zero, so this drift term reflects only waiting time effects. Holding workload in the initial buffers (buffers 1–2) generates a positive drift rate, because longer waiting times increase the likelihood that a customer requires an additional episode and transitions to a secondary buffer; see Table D.2. Holding workload in the secondary buffers (buffers 3–5) generates no such drift rate (in the base case), since customers cannot receive more than two episodes.

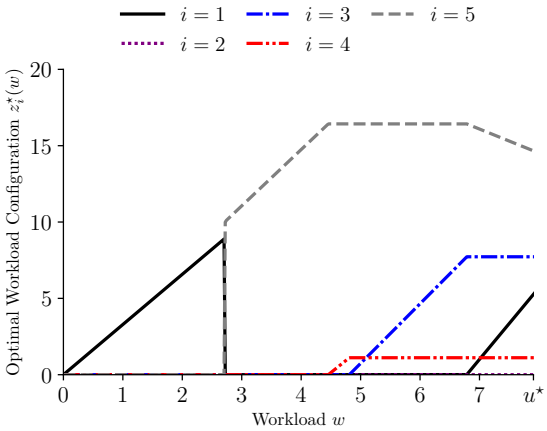
These two effects define the central trade-off governing the workload distribution. Holding workload in buffers with low holding cost h_i/m_i reduces immediate costs but increases future congestion, captured by the continuation cost $-d_i v^\star(w)$. The effective state cost per unit of work, $(h_i - d_i v^\star(w))/m_i$, aggregates the two costs, and the optimal workload configuration $z^\star(w)$ minimizes the total effective state cost; see (6.22).

Figure 8.2 makes this trade-off explicit: Figure 8.2a depicts the effective state costs per unit of work, which we hereafter call the effective costs. For the initial buffers (buffers 1–2), the effective cost increases with workload because longer waiting times increase the likelihood of a transition to a secondary buffer, and hence the continuation cost $-d_i v^\star(w)$. By contrast, for the secondary buffers (buffers 3–5), the effective cost is independent of workload, since customers cannot receive more than two service episodes.

Figure 8.2b depicts the optimal workload configuration. At low workload levels, class 1 has

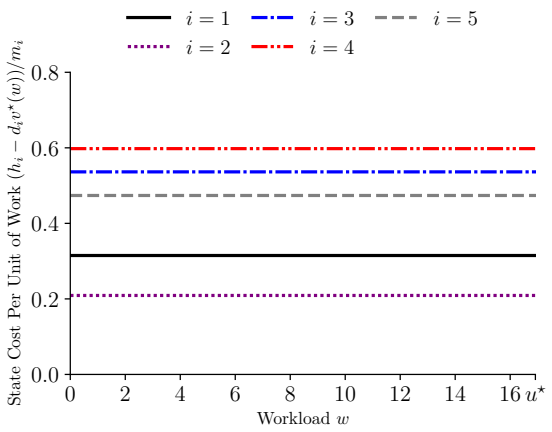


(a) Effective state cost per unit of work $(h_i - d_i v^\star(w))/m_i$.

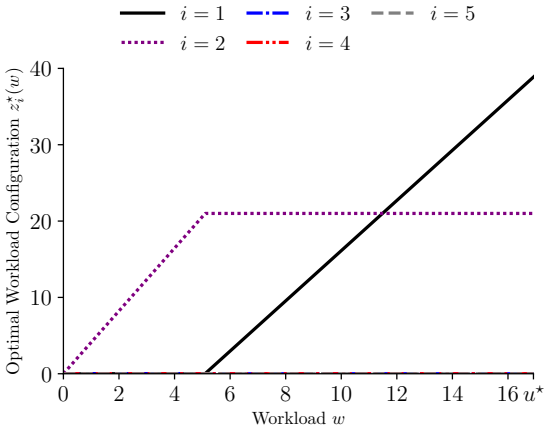


(b) Optimal workload configuration $z_i^\star(w)$.

Figure 8.2: Effective state cost per unit of work (effective cost) and optimal workload configuration for each class, as a function of the workload, in the base case ($\Delta_0 = 0$). The upper barrier is $u^\star = 7.91$.



(a) Effective state cost per unit of work $(h_i - d_i v^\star(w))/m_i$.



(b) Optimal workload configuration $z_i^\star(w)$.

Figure 8.3: Effective state cost per unit of work (effective cost) and optimal workload configuration for each class, as a function of the workload, in the base case ($\Delta_0 = 0$) when $c_{kj} = 0$ for $k = 1, 2$ and $j \in \mathcal{J}$ (i.e., waiting times do not affect service outcomes and transitions). The upper barrier is $u^\star = 16.92$.

the smallest effective cost, so the entire workload is held in buffer 1. As workload increases, the effective cost of buffer 1 increases (due to the continuation cost) until, at $w = 2.72$, it exceeds that of class 5, and workload is shifted to buffer 5. Buffer 5 fills at $w = 4.46$ (when the age of its head-of-line customer reaches the deadline), after which additional workload is allocated to buffer 4, and then, once buffer 4 fills at $w = 4.82$, to buffer 3 (in increasing order of effective cost). Once workload reaches $w = 6.79$, all secondary buffers (buffers 3–5) are full, and additional workload is again allocated to buffer 1. However, to avoid deadline violations, the workload allocated to buffer 5

is drawn down as more workload is allocated to buffer 1.[4] Thus, as congestion builds, the optimal workload allocation shifts from the class with the lowest immediate cost toward the classes that least amplify future congestion, subject to the age constraints.

Our proposed scheduling policy implements this workload allocation: it first prioritizes customers in danger of violating their deadlines, then allocates the remaining capacity to track the optimal workload configuration $z^\star(w)$, in the order induced by the effective costs.

To isolate the role of waiting time effects, Figure 8.3 depicts the effective costs and the resulting workload configuration when waiting times do not affect outcomes or transitions, i.e., when $c_{kj} = 0$ for $k = 1, 2$ and $j \in \mathcal{J}$. In this case, both the drift rate $-d_i z_i(w)$ and the continuation cost $-d_i v^\star(w)$ vanish, so the effective cost reduces to the immediate holding cost h_i/m_i, which is independent of workload. The resulting scheduling policy is therefore myopic: it holds the entire workload in the initial buffers at every workload level, ignoring the downstream congestion this creates; see Figure 8.3b. Section 8.4 shows that a policy built on this myopic workload configuration performs strikingly poorly in the original system.

Rejection Policy. When the workload reaches the upper barrier $u^\star = 7.91$, the marginal cost of holding an additional unit of work, $v^\star(w)$, exceeds the lowest rejection cost per unit of work. Thus, class 5 customers, which have the lowest rejection cost per unit of work, are rejected.

Service Rate Control Policy. Service rate adjustments affect the workload formulation through the effective drift rate $\theta(w)$. Setting $\theta(w) = x$ has two effects. First, it incurs an immediate cost at rate $\theta_0 x$, where $\theta_0 = \beta_j/(m'C)_j = 1.26$ for all basic activities $j = 1, \ldots, J_b$; see (5.7), (5.12), and Lemma 5.5. This term captures the aggregate cost in the original system associated with increasing service rates, such as higher transition costs due to poorer service outcomes. Second, the effective drift rate affects the evolution of the workload process. In particular, generating an effective drift rate x changes the expected future cost at rate $-xv^\star(w)$; see (5.13). The optimal effective drift rate $\theta^\star(w)$ balances immediate and future costs by minimizing $\theta_0 x - xv^\star(w)$, subject to admissible drift rate bounds implied by the maximum service rate variation constraint.

The result is a two-regime policy; see Figure D.2 (in Appendix D). When workload is low ($w < 0.96$, so $v^\star(w) < \theta_0$), the policy employs the lower service rate vector $100\mu^\star - 1.6$, prioritizing immediate service outcomes at the expense of future congestion. When workload is high ($w \geq 0.96$, so $v^\star(w) \geq \theta_0$), the policy employs the higher service rate vector $100\mu^\star + 1.6$, mitigating future congestion even though faster service can worsen outcomes for customers currently in service. Thus, service rates increase with congestion.

[4] A unit of workload generates less waiting in class 1 than in class 5 because buffer 1 has a higher effective output rate.

Table 8.2: Summary of the six policies considered in the simulation study.

Policy	Scheduling	Service Rate	Baseline Rate
Static Sch., Nominal Static Rates (S-NS)	Static	Nominal static	0
Static Sch., Opt. Static Rates (S-OS)	Static	Optimized static	Optimized
Static Sch., Dyn. Rates, Zero Base. (S-DZ)	Static	Dynamic	0
Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait (D-DZ-IW)†	Dynamic	Dynamic	0
Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	Dynamic	Dynamic	0
Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	Dynamic	Dynamic	Optimized

Notes. † Computed as if waiting times did not affect outcomes or transitions, i.e., $c_{kj} = 0$ for $k = 1, 2$ and $j \in \mathcal{J}$.

8.3 Description of Policies

To evaluate our proposed policy and quantify the value of dynamic scheduling and service rate control, we compare six policies, three with static scheduling and three with dynamic scheduling; see Table 8.2 for a summary. In all policies, the baseline service rate vector Δ_0 takes the form (6.24), with a common baseline rate for basic activities and zero for nonbasic activities. Each policy is simulated 100 times for 10^5 units of time, starting from an empty system. To eliminate the transient effects, the initial 20% of each run is discarded.

Static Scheduling Policies. All three use the static scheduling and rejection policies derived from the optimal solution to the SPP (4.2) and differ only in the service rate control policy. The first policy, called static scheduling with nominal static service rates (S-NS), employs the nominal service rate vector $r\mu^{\star}$. The second policy, called static scheduling with optimized static service rates (S-OS), employs the static service rate vector $\mu^r = r\mu^{\star} + \sqrt{r}\Delta_0$, where Δ_0 is optimized by simulation (to a precision of 0.01) to minimize the long-run average cost. The third policy, called static scheduling with dynamic service rates and zero baseline rates (S-DZ), employs the dynamic service rates prescribed by our proposed policy for $\Delta_0 = 0$; see (7.1).

Dynamic Scheduling Policies. The three dynamic scheduling policies differ in whether they account for waiting time effects and whether the baseline service rate is optimized. The first policy, called dynamic scheduling with dynamic service rates and zero baseline rates, ignoring waiting time effects (D-DZ-IW), uses the proposed policy with $\Delta_0 = 0$, computed as if waiting times did not affect outcomes or transitions (i.e., $c_{kj} = 0$ for $k = 1, 2$ and $j \in \mathcal{J}$). The second policy, called dynamic scheduling with dynamic service rates and zero baseline rates (D-DZ), uses the proposed policy with $\Delta_0 = 0$. The third policy, called dynamic scheduling with dynamic service rates and optimized baseline rates (D-DO), uses the proposed policy in which Δ_0 is optimized as in the S-OS policy.

Table 8.3: Comparative evaluation of the long-run average cost of the various policies.

Policy	Maximum Service Rate Variation (Δ)				
	0.00	0.08	0.16	0.24	0.32
Static Sch., Nominal Static Rates (S-NS)	182.98 ± 0.44	–	–	–	–
Static Sch., Opt. Static Rates (S-OS)	38.45 ± 0.06	–	–	–	–
Static Sch., Dyn. Rates, Zero Base. (S-DZ)	182.98 ± 0.44	82.46 ± 0.35	45.01 ± 0.15	35.24 ± 0.08	32.82 ± 0.05
Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	274.77 ± 0.27	240.75 ± 0.29	207.00 ± 0.29	172.27 ± 0.29	138.63 ± 0.31
Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	49.61 ± 0.11	39.87± 0.11	32.03 ± 0.08	27.44 ± 0.07	25.47 ± 0.05
Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	32.60 ± 0.06	29.67 ± 0.06	27.40 ± 0.05	25.92 ± 0.05	25.22 ± 0.05

Notes. The reported values are sample averages and their associated 95% confidence intervals.

8.4 Results and Managerial Insights

Table 8.3 reports the long-run average cost of the various policies for the base case; The corresponding results for deadlines of one and three time units are reported in Table D.4 (in Appendix D). Three patterns emerge. First, the substantial gap between S-NS and S-OS highlights the suboptimality of relying solely on the SPP (4.2) to determine service rates. While the nominal rates provide a first-order approximation, second-order adjustments are needed. Second, S-DZ improves substantially on S-NS, showing that adjusting service rates in response to congestion is valuable even when scheduling is held static. Third, the dynamic scheduling policies D-DZ and D-DO uniformly dominate their static counterparts, often by wide margins, demonstrating the importance of adjusting scheduling in response to congestion.

Table D.4 also highlights the interplay between scheduling and service rate flexibility. As the maximum service rate variation increases, the gap between the static scheduling policy S-DZ and the dynamic scheduling policies D-DZ and D-DO narrows, suggesting that dynamic scheduling and dynamic service rate control are partial substitutes. Nevertheless, even at the largest service rate variation, D-DZ and D-DO outperform S-DZ, indicating that dynamic service rate control does not fully substitute for dynamic scheduling. Moreover, although larger service rate adjustments can improve performance, the gains exhibit diminishing returns. Finally, optimizing baseline service rates (D-DO relative to D-DZ) helps only when service rate flexibility is low; with greater flexibility, dynamic adjustment compensates for zero baseline rates.

Accounting for Waiting Time Effects. The D-DZ-IW policy, derived as if waiting times did not affect outcomes or transitions, isolates the cost of ignoring waiting time effects. As discussed in Section 8.2, ignoring waiting time effects results in a myopic policy that bases its decisions solely on immediate costs and concentrates the entire workload in the initial buffers (buffers 1–2), even at high workload levels; see Figure 8.3b. Table D.4 demonstrates that this workload misallocation leads to substantial underperformance. In particular, D-DZ-IW is uniformly the worst-performing dynamic scheduling policy and its performance degrades as deadlines increase because longer deadlines let

customers linger and amplify the cost of a myopic workload distribution. Most strikingly, except at high service rate flexibility levels, D-DZ-IW is dominated even by the static scheduling policies, underscoring the importance of accounting for waiting time effects.

Dynamic Service Rates. Table D.9 (in Appendix D) reports the optimized baseline service rates (Δ_0) for the S-OS and D-DO policies. When service rate flexibility is low, baseline service rates are positive, indicating the need for service rates higher than the nominal service rates to mitigate congestion. For instance, with a deadline of two time units and maximum service rate variation $\Delta = 0$, D-DO employs the service rate vector $100\mu^\star + 1.1$ in all workload levels; see Table 8.1 for the nominal service rates. As service rate flexibility increases, the optimized baseline service rates decrease. For instance, with $\Delta = 0.32$, D-DO employs the service rate vector $100\mu^\star - 1.3$ when $w < 1.0$ and $100\mu^\star + 1.9$ when $w \geq 1.0$, thereby reducing the long-run average cost by 23% relative to the same policy with $\Delta = 0$; see Table D.4.

As discussed in Section 8.2, when congestion is high, higher service rates are required to reduce future congestion costs, even though they may worsen outcomes for customers currently in service. Conversely, when workload is low, lower service rates, sometimes below the nominal rates, are required to improve outcomes even though they increase future congestion costs. That is, in multi-episode service systems with waiting time-dependent and stochastic transitions, higher congestion does not call for lower service rates solely to equalize outcomes across workload levels, contrary to conventional intuition.

Figures D.3–D.4 (in Appendix D) illustrate the interplay between dynamic service rate control and dynamic scheduling. They depict the effective costs and the resulting workload configuration under D-DO with $\Delta = 0$ and $\Delta = 0.32$, respectively. With greater service rate flexibility ($\Delta = 0.32$), D-DO employs higher service rates at high workload levels ($w \geq 1.0$) to mitigate congestion, which reduces the continuation costs $d_i v^\star(w)$ associated with the primary buffers (buffers 1–2). Consequently, workload is held (entirely) in buffer 1 over a wider range, for $w < 2.82$ with $\Delta = 0.32$ (Figure D.4b) versus $w < 2.66$ with $\Delta = 0$ (Figure D.3b). Thus, greater service rate flexibility triggers a shift to secondary buffers at larger workloads.

Waiting Time Sensitivity. To examine the impact of the waiting time sensitivity of outcomes and transitions, we vary c_{11} while holding the ratios c_{1j}/c_{11} and c_{2j}/c_{11} fixed for $j \in \mathcal{J}$; see Tables D.5 and D.10 (in Appendix D). As waiting time sensitivity increases, the performance of all policies deteriorates because congestion generates a larger positive drift rate in the workload process. Moreover, the gap between S-NS and policies with dynamic control widens, underscoring the value of dynamic control under high waiting time sensitivity. Furthermore, the optimized baseline service rates (Δ_0) increase to generate a larger negative drift rate and offset the larger congestion-driven positive drift rate induced by waiting times; see Table D.10.

Figures D.5–D.6 (in Appendix D) illustrate how waiting time sensitivity shapes the scheduling policy. They depict the effective costs and the resulting workload configuration for sensitivity levels equal to one-half and twice those in the base case. As waiting time sensitivity increases, holding workload in the initial buffers (buffers 1–2) generates a larger positive drift rate, so the continuation costs $-d_i v^\star(w)$ rise more steeply with workload. Consequently, workload is held (entirely) in buffer 1 over a narrower range of workloads, shrinking from $w < 8.49$ under low sensitivity (Figure D.5b) to $w < 0.58$ under high sensitivity (Figure D.6b). Thus, higher sensitivity triggers the shift to secondary buffers at lower workload levels.

Abandonment Rates. To examine the impact of abandonment rates, we vary ℓ_1 while holding $\ell_i = \ell_1$ for $i \in \mathcal{I}$; see Tables D.6 and D.11 (in Appendix D). As abandonment rates increase, the performance of all policies improves. Abandonment is costly in isolation (its cost exceeds the nominal service cost, though not the rejection cost, in this example; see Tables 8.1 and D.2), but it generates a negative drift rate that partially offsets the positive drift rate induced by waiting times, reducing congestion and total costs. Figures D.7–D.8 (in Appendix D) show that abandonment can alter the scheduling policy. As abandonment rates increase, the continuation costs $-d_i v^\star(w)$ decrease across all buffers (see also (4.14) and (5.3)), altering the ordering of the effective costs. When $\ell_1 = 0.03$, workload is first allocated to buffer 1 (Figure D.7b). However, when $\ell_1 = 0.1$, workload is first allocated to secondary buffers, and then buffer 1 (Figure D.8b).

Abandonment Costs. To examine the impact of abandonment costs, we fix $\ell_i = 0.03$ for $i \in \mathcal{I}$ and vary π_1 while holding the ratios $\pi_i/\nu_i = \pi_1/\nu_1$ fixed for $i \in \mathcal{I}$; see Tables D.7 and D.12 (in Appendix D). As abandonment costs increase, the performance of all static scheduling policies and the dynamic scheduling policies D-DZ and D-DO deteriorates, as expected. However, the gap between D-DZ-IW and D-DZ narrows. Because abandonment costs constitute a larger share of the effective costs $(h_i - d_i v^\star(w))/m_i$ through their impact on h_i (see (4.15) and (5.10)), which reduces the relative importance of continuation costs, and hence waiting time effects. Thus, workload configuration of the myopic policy D-DZ-IW approaches that of D-DZ.

Waiting Costs. To examine the impact of waiting costs, we set $c_{3j} = c_{31}$ for $j \in \mathcal{J}$ and vary c_{31}; see Tables D.8 and D.13 (in Appendix D). Across all waiting cost levels, policies that employ dynamic scheduling and service rate control outperform their static counterparts, indicating that responsiveness to congestion remains valuable even when waiting costs are high. As waiting costs increase, S-OS and D-DO employ higher optimized baseline service rates (Δ_0) to reduce waiting times. Nonetheless, when service rate flexibility is high, both policies still employ service rates below the nominal rates at low workloads.

Figures D.11–D.12 (in Appendix D) illustrate how waiting costs shape the scheduling policy. Higher waiting costs increase the immediate holding cost component of the effective cost, so the

Table 8.4: Comparative evaluation of the long-run average costs in the discrete, non-preemptive setting.

Policy	Maximum Service Rate Variation (Δ)				
	0.00	0.08	0.16	0.24	0.32
Static Sch., Nominal Static Rates (S-NS)	283.47 ± 0.15	–	–	–	–
Static Sch., Opt. Static Rates (S-OS)	48.09 ± 0.05	–	–	–	–
Static Sch., Dyn. Rates, Zero Base.(S-DZ)	283.47 ± 0.15	255.80 ± 0.13	229.06 ± 0.16	201.08 ± 0.15	161.44 ± 0.30
Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	420.09 ± 0.27	383.02 ± 0.25	346.10 ± 0.24	309.41 ± 0.24	271.67 ± 0.24
Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	58.48 ± 0.13	41.79 ± 0.12	28.28 ± 0.09	22.09 ± 0.07	20.24 ± 0.04
Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	28.43 ± 0.07	24.63 ± 0.05	22.06 ± 0.04	20.55 ± 0.04	20.07 ± 0.04

Notes. The reported values are sample averages and their associated 95% confidence intervals.

continuation cost $-d_1 v^\star(w)$ of buffer 1 must rise further before it outweighs the higher holding cost of the secondary buffers. Consequently, workload is held (entirely) in buffer 1 over a wider range, with $w < 5.81$ under high waiting costs (Figure D.12b) compared to $w < 0.84$ under low waiting costs (Figure D.11b). Thus, higher waiting costs trigger the shift to secondary buffers at higher workloads.

8.5 A Discrete, Non-Preemptive Implementation

In rehabilitation centers, service is delivered through discrete appointments, a patient is told at the outset how many appointments they will receive in each episode, and service is not interrupted mid-episode. Our analysis thus far abstracts away from these features. To assess the robustness of our policy to these features, this section considers a discrete-time, non-preemptive setting in which a patient's service rate is held fixed over each episode. To be specific, patients receive two appointments per week, so we divide each 50-week year into 100 periods. In each period, every patient on a clinician's caseload receives one appointment. At the beginning of each period, the system manager applies the scheduling, service rate, and rejection control policies of Chapter 7, with the exception that when an episode begins, the number of appointments the patient will receive in this episode is determined by the current service rate, and the patient remains in service, without interruption, until all are delivered. All other primitives are as in the base case of Section 8.1.

The findings of Section 8.4 carry over to this setting; see Table 8.4 for the results in the base case and Tables D.14–D.19 (in Appendix D) for the full set of results. The dynamic scheduling policies D-DZ and D-DO uniformly dominate their static counterparts, often by wide margins. For instance, with a deadline of two units of time and maximum service rate variation $\Delta = 0.32$, D-DO attains a long-run average cost of 20.07, compared with 48.09 under S-OS and 161.44 under S-DZ; see Table D.14. D-DZ-IW continues to perform poorly across all parameter regimes, reaffirming that accounting for waiting time effects is essential. The comparative statics also match

our earlier findings: longer deadlines and higher waiting time sensitivity sharply degrade the static and D-DZ-IW policies while leaving D-DZ and D-DO largely unaffected; higher waiting time sensitivity calls for higher optimized baseline service rates under D-DO; and higher abandonment rates decrease costs and narrow the gap between D-DZ-IW and D-DZ. Thus, the structural insights of Section 8.4 carry over to the discrete, non-preemptive setting.

These results are not coincidental. The proposed policy responds to the aggregate workload, whose drift is driven by abandonments and waiting time effects. The workload process evolves on a slower timescale than individual arrivals and service completions, whereas the features considered here act only at the level of individual service episodes. Thus, they leave the workload dynamics and the policy structure unchanged.

Chapter 9

Conclusion

Motivated by pediatric rehabilitation, this thesis has studied multi-episode service systems in which customers may need several episodes to achieve the desired outcome, and in which each episode's outcome and the subsequent transition depend jointly on the service rate and the waiting time. Congestion is self-reinforcing: longer waiting times worsen outcomes and create downstream demand, so actions that lower immediate costs can amplify future congestion. We model the system as a stochastic processing network and analyze the resulting control problem in heavy traffic. We show that under complete resource pooling, this problem can be approximated by a one-dimensional workload formulation whose optimal solution is a two-sided barrier policy with a state-dependent drift rate. Building on this solution, we propose an easily implementable, tractable dynamic scheduling, service rate, and rejection control policy for the original service system.

Our analysis yields several structural insights. First, explicitly accounting for the impact of waiting times on outcomes and transitions is essential; policies that ignore it can perform poorly, particularly as waiting time sensitivity increases. Second, scheduling and service rate control both affect the evolution of congestion, but through distinct mechanisms. Service rate control regulates congestion directly, while scheduling shapes it indirectly by determining who waits. Policies that jointly exploit both levers consistently outperform those relying on only one. Third, scheduling priorities are driven by effective state costs that balance immediate congestion costs against the effect of waiting times on future congestion, subject to age constraints. Under low congestion, immediate costs dominate and high-immediate-cost classes take priority; as congestion rises, priority shifts to the classes whose waiting most amplifies downstream congestion. Fourth, service rate control exhibits a regime-switching structure: under low congestion, it favors outcome improvement at the cost of future congestion, whereas under high congestion, it prioritizes congestion mitigation at the expense of current outcomes. Service rates thus rise with congestion, contrary to the conventional view that they should fall to equalize outcomes; in multi-episode systems with stochastic, waiting

time dependent outcomes, such slowdown policies can exacerbate congestion and degrade long-run performance. Finally, abandonment, though costly, generates a negative drift that partially offsets the positive drift induced by waiting times, making it a partial substitute for service rate control. Moreover, since abandonment alters the evolution of congestion, it interacts closely with scheduling decisions and influences who bears waiting.

The analysis also yields concrete operational prescriptions. When the system is congested, deliver (slightly) fewer appointments per episode, prioritize classes by both the downstream congestion they create and their immediate cost, and reject only as a last resort. Since the policy tracks the aggregate workload, it primarily responds to coarse changes in congestion, and requires small adjustments to appointment counts.

References

Hossein Abouee-Mehrizi, Nasser Barjesteh, Benjamin Ravencroft, Armita Tehranchi, and Brendan Wylie-Toal. The impact of treatment modality and waiting time: An empirical analysis of a pediatric speech-language therapy program. *Working Paper*, 2026.

Philipp Afèche. Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manuf. Serv. Oper. Manag.*, 15(3):423–443, 2013.

Bahram Alidaee and Norman Keith Womer. Scheduling with time dependent processing times: review and extensions. *J. Oper. Res. Soc.*, 50(7):711–720, 1999.

Amir Anastasios Alwan and Nasser Barjesteh. Dynamic pricing, scheduling, and order rejection of a multiclass omnichannel hybrid production system. *Working Paper*, 2025.

Amir Anastasios Alwan, Baris Ata, and Yuwei Zhou. A queueing model of dynamic pricing and dispatch control for ride-hailing systems incorporating travel times. *Queueing Systems*, 2024.

Baris Ata. Dynamic control of a multiclass queue with thin arrival streams. *Oper. Res.*, 54(5): 876–892, 2006.

Baris Ata and Nasser Barjesteh. An approximate analysis of dynamic pricing, outsourcing, and scheduling policies for a multiclass make-to-stock queue in the heavy traffic regime. *Oper. Res.*, 71(1):341–357, 2022.

Baris Ata and Sunil Kumar. Heavy traffic analysis of open processing networks with complete resource pooling: Asymptotic optimality of discrete review policies. *Ann. Appl. Probab.*, 15(1A): 331–391, 2005.

Baris Ata and Tava Lennon Olsen. Near-optimal dynamic lead-time quotation and scheduling under convex-concave customer delay costs. *Oper. Res.*, 57(3):753–768, 2009.

Baris Ata and Mustafa H Tongarlak. On scheduling a multiclass queue with abandonments under general delay costs. *Queueing Systems*, 74(1):65–104, 2013.

Baris Ata, J Michael Harrison, and Lawrence A Shepp. Drift rate control of a Brownian processing system. *Ann. Appl. Probab.*, 15(2):1145–1160, 2005.

Baris Ata, Deishin Lee, and Erkut Sönmez. Dynamic volunteer staffing in multicrop gleaning operations. *Oper. Res.*, 67(2):295–314, 2019.

Baris Ata, Nasser Barjesteh, and Sunil Kumar. Dynamic dispatch and centralized relocation of cars in ride-hailing platforms. *Working Paper*, 2020.

Baris Ata, Deishin Lee, and Mustafa Hayri Tongarlak. A diffusion model of dynamic participant inflow management. *Queueing Systems*, pages 1–32, 2024a.

Baris Ata, Mustafa H Tongarlak, Deishin Lee, and Joy Field. A dynamic model for managing volunteer engagement. *Oper. Res.*, 2024b.

Achal Bassamboo and Ramandeep Singh Randhawa. Scheduling homogeneous impatient customers. *Management Sci.*, 62(7):2129–2147, 2015.

René Bekker, Onno Johan Boxma, and Jacobus Adrianus Cornelis Resing. Lévy processes with adaptable exponent. *Adv. Appl. Probab.*, 41(1):177–205, 2009.

Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena scientific, 1997.

Onno J Boxma and Maria Vlasiou. On queues with service and interarrival times depending on waiting times. *Queueing Systems*, 56(3):121–132, 2007.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

PH Brill and MJM Posner. A two server queue with nonwaiting customers receiving specialized service. *Management Sci.*, 27(8):914–925, 1981.

Sid Browne and Uri Yechiali. Scheduling deteriorating jobs on a single processor. *Oper. Res.*, 38(3): 495–498, 1990.

Amarjit Budhiraja, Arka P Ghosh, and Chihoon Lee. Ergodic rate control problem for single class queueing networks. *SIAM J. Control Optim.*, 49(4):1570–1606, 2011.

Sabri Çelik and Costis Maglaras. Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. *Management Sci.*, 54(6):1132–1146, 2008.

Donald B Chalfin, Stephen Trzeciak, Antonios Likourezos, Brigitte M Baumann, R Phillip Dellinger, and Delay-Ed Study Group. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Crit. Care Med.*, 35(6):1477–1483, 2007.

Carri W Chan, Vivek F Farias, Nicholas Bambos, and Gabriel J Escobar. Optimizing intensive care unit discharge decisions with patient readmissions. *Oper. Res.*, 60(6):1323–1341, 2012.

Carri W Chan, Galit Yom-Tov, and Gabriel Escobar. When to use speedup: An examination of service systems with returns. *Oper. Res.*, 62(2):462–482, 2014.

Carri W Chan, Vivek F Farias, and Gabriel J Escobar. The impact of delays on service times in the intensive care unit. *Management Sci.*, 63(7):2049–2072, 2017.

Paul S Chan, Harlan M Krumholz, Graham Nichol, Brahmajee K Nallamothu, and American Heart Association National Registry of Cardiopulmonary Resuscitation Investigators. Delayed time to

defibrillation after in-hospital cardiac arrest. *N. Engl. J. Med.*, 358(1):9–17, 2008.

TC Edwin Cheng, Qing Ding, and Bertrand MT Lin. A concise survey of scheduling with time-dependent processing times. *Eur. J. Oper. Res.*, 152(1):1–13, 2004.

Children's Hospital of Richmond. New therapy scheduling process improves access and empowers patients and families. https://www.chrichmond.org/blog/details/new-therapy-scheduling-process-improves-access-and-empowers-patients-and-families, 2023. (Accessed: 2026-01-01).

Earl A. Coddington and Norman Levinson. *Theory of Ordinary Differential Equations*. McGraw-Hill, 1955.

Barbara Jane Cunningham, Karla N Washington, Amanda Binns, Katelyn Rolfe, Bernadette Robertson, and Peter Rosenbaum. Current methods of evaluating speech-language outcomes for preschoolers with communication disorders: A scoping review using the ICF-CY. *J. Speech Lang. Hear. Res.*, 60(2):447–464, 2017.

Jim G Dai and Wuqin Lin. Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Ann. Appl. Probab.*, 18(6):2239–2299, 2008.

Bernardo D'Auria, Ivo JBF Adan, René Bekker, and Vidyadhar Kulkarni. An M/M/c queue with queueing-time dependent service rates. *Eur. J. Oper. Res.*, 299(2):566–579, 2022.

Jelle De Vries, Debjit Roy, and René De Koster. Worth the wait? how restaurant waiting time influences customer behavior and revenue. *J. Oper. Manag.*, 63:59–78, 2018.

Jing Dong, Pnina Feldman, and Galit B Yom-Tov. Service systems with slowdowns: Potential failures and proposed solutions. *Oper. Res.*, 63(2):305–324, 2015.

Jing Dong, Berk Gorgulu, and Vahid Sarhangian. Multiclass queue scheduling under slowdown: An approximate dynamic programming approach. *Working Paper*, 2025.

Uwe Einmahl. Strong invariance principles for partial sums of independent random vectors. *Ann. Probab.*, pages 1419–1440, 1987a.

Uwe Einmahl. A useful estimate in the multidimensional invariance principle. *Probab. Theory Relat. Fields*, 76(1):81–101, 1987b.

Uwe Einmahl. Extensions of results of Komlós, Major, and Tusnády to the multivariate case. *J. Multivar. Anal.*, 28(1):20–68, 1989.

Stewart N Ethier and Thomas G Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, 2nd edition, 2009.

William Feller. *An introduction to probability theory and its applications, Volume 2*, volume 81. John Wiley & Sons, 1991.

Xuefeng Gao and Junfei Huang. Asymptotically optimal control of make-to-stock systems. *Math.*

Oper. Res., 49(2):948–985, 2023.

Samim Ghamami and Amy R Ward. Dynamic scheduling of a two-server parallel server system with complete resource pooling and reneging in heavy traffic: Asymptotic optimality of a two-threshold policy. *Math. Oper. Res.*, 38(4):761–824, 2013.

Arka P Ghosh and Ananda P Weerasinghe. Optimal buffer size for a stochastic processing network in heavy traffic. *Queueing Systems*, 55(3):147–159, 2007.

Arka P Ghosh and Ananda P Weerasinghe. Optimal buffer size and dynamic rate control for a queueing system with impatient customers in heavy traffic. *Stoch. Process. Their Appl.*, 120(11): 2103–2141, 2010.

Katherine E Harding, Annie K Lewis, Amy Dennett, Kylie Hughes, Michelle Clarke, and Nicholas F Taylor. An evidence-based demand management strategy using a hub and spoke training model reduces waiting time for children's therapy services: An implementation trial. *Child Care Health Dev.*, 50(1):e13154, 2024.

J Michael Harrison. Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic Differential Systems, Stochastic Control Theory and Applications*, pages 147–186. Springer, 1988.

J Michael Harrison. Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete-review policies. *Ann. Appl. Probab.*, pages 822–848, 1998.

J Michael Harrison. Brownian models of open processing networks: Canonical representation of workload. *Ann. Appl. Probab.*, 10(1):75–103, 2000.

J Michael Harrison. A broader view of Brownian networks. *Ann. Appl. Probab.*, 13(3):1119–1150, 2003.

J Michael Harrison. *Brownian Models of Performance and Control*. Cambridge University Press, 2013.

J Michael Harrison and Marcel J López. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems*, 33(4):339–368, 1999.

J Michael Harrison and Vien Nguyen. The qnet method for two-moment analysis of open queueing networks. *Queueing systems*, 6:1–32, 1990.

J Michael Harrison and Vien Nguyen. Brownian models of multiclass queueing networks: Current status and open problems. *Queueing Systems*, 13:5–40, 1993.

J Michael Harrison and Lawrence M Wein. Scheduling networks of queues: Heavy traffic analysis of a simple open network. *Queueing Systems*, 5(4):265–279, 1989.

J Michael Harrison and Lawrence M Wein. Scheduling networks of queues: Heavy traffic analysis of a two-station closed network. *Oper. Res.*, 38(6):1052–1064, 1990.

J Michael Harrison and Ruth J. Williams. Workload reduction of a generalized Brownian network. *Ann. Appl. Probab.*, 15(4):2255–2295, 2005.

J Michael Harrison and Ruth J. Williams. Workload interpretation for Brownian models of stochastic processing networks. *Math. Oper. Res.*, 32(4):808–820, 2007.

Yue Hu, Carri W Chan, and Jing Dong. Optimal scheduling of proactive service with customer deterioration and improvement. *Management Sci.*, 68(4):2533–2578, 2022.

Junfei Huang, Boaz Carmeli, and Avishai Mandelbaum. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Oper. Res.*, 63(4):892–908, 2015.

Qing Huang, Amardeep Thind, Jonathan F Dreyer, and Gregory S Zaric. The impact of delays to admission from the emergency department on inpatient outcomes. *BMC Emergency Medicine*, 10 (1):16, 2010.

Rouba Ibrahim and Ward Whitt. Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Sci.*, 55(10):1729–1742, 2009.

Rouba Ibrahim, Mor Armony, and Achal Bassamboo. Does the past predict the future? The case of delay announcements in service systems. *Management Sci.*, 63(6):1762–1780, 2017.

Karl G Kempf, Pınar Keskinocak, and Reha Uzsoy. *Planning production and inventories in the extended enterprise: a state of the art handbook*, volume 1. Springer Science & Business Media, 2011.

KidsAbility. Preschool speech and language services. https://www.kidsability.ca/programs-services/preschool-speech-and-language-services/, 2026. (Accessed: 2026-01-01).

Jeunghyun Kim, Ramandeep S Randhawa, and Amy R Ward. Dynamic scheduling in a many-server, multiclass system: The role of customer impatience in large systems. *Manuf. Serv. Oper. Manag.*, 20(2):285–301, 2018.

Elena V Krichagina and Michael I Taksar. Diffusion approximation for GI/G/1 controlled queues. *Queueing Systems*, 12(3-4):333–367, 1992.

Elena V Krichagina, Sheldon XC Lou, Suresh P Sethi, and Michael I Taksar. Production control in a failure-prone manufacturing system: Diffusion approximation and asymptotic optimality. *Ann. Appl. Probab.*, 3(2):421–453, 1993.

Sunil Kumar. Two-server closed networks in heavy traffic: Diffusion limits and asymptotic optimality. *Ann. Appl. Probab.*, 10(3):930–961, 2000.

Xiaole Alyssa Liu, Jing Dong, Yosef Berlyand, and Martin Copenhaver. Measuring the heterogeneous effect of emergency department boarding on inpatient length of stay: Toward efficient and equitable inpatient bed assignment. *Working Paper*, 2025.

Lurie Children's Hospital. Speech-language pathology rehabilitation services: Mod-

els of care. https://www.luriechildrens.org/globalassets/media/pages/specialties--conditions/specialities/speech-language-pathology/rehab-services-models-of-care-flyer-english.pdf, 2023. (Accessed: 2026-01-01).

Avi Mandelbaum and Gennady Pats. State-dependent stochastic networks. Part I. approximations and applications with continuous diffusion limits. *Ann. Appl. Probab.*, 8(2):569–646, 1998.

Nicole McGill, Kathryn Crowe, and Sharynne Mcleod. "many wasted months": Stakeholders' perspectives about waiting for speech-language pathology services. *Int. J. Speech Lang. Pathol.*, 22(3):313–326, 2020.

Zbigniew Palmowski and Maria Vlasiou. A lévy input model with additional state-dependent services. *Stoch. Process. Appl.*, 121(7):1546–1564, 2011.

Erica Plambeck, Sunil Kumar, and J Michael Harrison. A multiclass queue in heavy traffic with throughput time constraints: Asymptotically optimal dynamic controls. *Queueing Systems*, 39(1): 23–54, 2001.

Erica L Plambeck. Optimal leadtime differentiation via diffusion approximations. *Oper. Res.*, 52(2): 213–228, 2004.

M Posner. Single-server queues with service time dependent on waiting time. *Oper. Res.*, 21(2): 610–616, 1973.

JE Reed and Amy R Ward. Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Math. Oper. Res.*, 33(3):606–644, 2008.

Michael Reich, Avi Mandelbaum, and Ya'acov Ritov. The workload process: Modeling, inference and applications. 2010.

Martin I Reiman. Open queueing networks in heavy traffic. *Math. Oper. Res.*, 9(3):441–458, 1984.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 2015.

Melanie Rubino and Baris Ata. Dynamic control of a make-to-order, parallel-server system with cancellations. *Oper. Res.*, 57(1):94–108, 2009.

Jori Selen, Ivo JBF Adan, Vidyadhar G Kulkarni, and Johan SH van Leeuwaarden. The snowball effect of customer slowdown in critical many-server systems. *Stoch. Models*, 32(3):366–391, 2016.

Pengyi Shi, Jonathan E Helm, Jivan Deglise-Hawkinson, and Julian Pan. Timing it right: Balancing inpatient congestion vs. readmission risk at discharge. *Oper. Res.*, 69(6):1842–1865, 2021.

AW Siegmeth, K Gurusamy, and MJ Parker. Delay to surgery prolongs hospital stay in patients with fractures of the proximal femur. *J. Bone Joint Surg. Br.*, 87(8):1123–1126, 2005.

Shaler Stidham. Scheduling, routing, and flow control in stochastic networks. In *Stochastic Differential Systems, Stochastic Control Theory and Applications*, pages 529–561. Springer, 1988.

Shaler Stidham. Analysis, design, and control of queueing systems. *Oper. Res.*, 50(1):197–216, 2002.

Huifeng Su, Lesley Meng, Rohit Sangal, and Edieal J Pinker. Emergency department boarding: Quantifying the impact of ed boarding on patient outcomes and downstream hospital operations. *Working Paper*, 2024.

Xu Sun and Xiaohan Zhu. Dynamic control of a make-to-order system under model uncertainty. *Management Sci.*, 2024.

Rangarajan K Sundaram. *A first course in optimization theory*. Cambridge University Press, 1996.

Amy R Ward and Sunil Kumar. Asymptotically optimal admission control of a queue with impatient customers. *Math. Oper. Res.*, 33(1):167–202, 2008.

Ananda Weerasinghe. Diffusion approximations for G/M/n+GI queues with state-dependent service rates. *Math. Oper. Res.*, 39(1):207–228, 2014.

Lawrence M Wein. Dynamic scheduling of a multiclass make-to-stock queue. *Oper. Res.*, 40(4): 724–735, 1992.

Ward Whitt. Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems*, 6(1):335–351, 1990.

Ward Whitt. Proofs of the martingale FCLT. *Probab. Surveys*, 4(4):268 – 302, 2007.

Chenguang Wu, Achal Bassamboo, and Ohad Perry. Service system with dependent service and patience times. *Management Sci.*, 65(3):1151–1172, 2019.

Chenguang Wu, Achal Bassamboo, and Ohad Perry. When service times depend on customers' delays: A relationship between two models of dependence. *Oper. Res.*, 70(6):3345–3354, 2022.

Galit B Yom-Tov and Carri W Chan. Balancing admission control, speedup, and waiting in service systems. *Queueing systems*, 97(1):163–219, 2021.

Appendix A

Supplementary Material for the Model

A.1 Characterization of the Cumulative Routing Processes

This section presents a construction of the cumulative routing processes. Let $\mathfrak{X}_l^j(m)$ be a binary random variable that equals one if the m-th customer served by activity j subsequently transitions to class l and equals zero otherwise for $j \in \mathcal{J}$, $l \in \mathcal{I}$, and $m \in \mathbb{N}$. Letting $\mathfrak{X}^j(m) = (\mathfrak{X}_l^j(m))$, we refer to $\{\mathfrak{X}^j(m) : m \in \mathbb{N}\}$ as the sequence of routing indicators for activity j for $j \in \mathcal{J}$. For $j \in \mathcal{J}$ and $m \in \mathbb{N}$, we let

$$\phi^j(m) \coloneqq e^{i(j)} - \mathfrak{X}^j(m), \tag{A.1}$$

where $e^{i(j)}$ is the unit vector with a one in the $i(j)$-th component. Then, the partial sum

$$\Phi_l^j(m) \coloneqq \phi_l^j(1) + \cdots + \phi_l^j(m) \tag{A.2}$$

represents the cumulative decrease in the number of class l customers as a result of the first m activity j service completions for $j \in \mathcal{J}$, $l \in \mathcal{I}$, and $m \in \mathbb{N}$. Letting $\Phi^j(m) = (\Phi_l^j(m))$, we define the cumulative routing process associated with activity j as $\Phi^j = \{\Phi^j(m) : m \in \mathbb{N}\}$ for $j \in \mathcal{J}$.

A.2 An Alternative Characterization of the Cumulative Routing Processes

This section presents an alternative characterization of the cumulative routing processes, which will prove useful in the derivation of the approximating Brownian control problem in Appendix B. Recall

from Appendix A.1 that $\mathfrak{X}_l^j(m)$ is a binary random variable that equals one if the m-th customer served by activity j subsequently transitions to class l and equals zero otherwise for $j \in \mathcal{J}$, $l \in \mathcal{I}$, and $m \in \mathbb{N}$. Letting t_m^j denote the arrival time of the m-th customer served by activity j, we have that

$$\mathbb{P}\big(\mathfrak{X}_l^j(m) = 1\big) \;=\; P_{jl}(\mu_j(t_m^j + w_{i(j)}(t_m^j)), w_{i(j)}(t_m^j))$$

for $j \in \mathcal{J}$, $l \in \mathcal{I}$, and $m \in \mathbb{N}$. For notational convenience, we define

$$\Delta P_{jlm} \;\coloneqq\; P_{jl}\big(\mu_j(t_m^j + w_{i(j)}(t_m^j)), w_{i(j)}(t_m^j)\big) - P_{jl}(\mu_j^\star, 0)$$

for $j \in \mathcal{J}$, $l \in \mathcal{I}$, and $m \in \mathbb{N}$. We begin by decomposing the sequence of routing indicators to isolate nominal routing dynamics from the effects of dynamic service rate adjustments and system congestion. To be specific, for $j \in \mathcal{J}$ and $l \in \mathcal{I}$, we write the sequence $\{\mathfrak{X}_l^j(m) : m \in \mathbb{N}\}$ as the sum of two sequences: $\{\bar{\mathfrak{X}}_l^j(m) : m \in \mathbb{N}\}$, which is a sequence of independent and identically distributed Bernoulli random variables with probability of success $P_{jl}(\mu_j^\star, 0)$, and $\{\hat{\mathfrak{X}}_l^j(m) : m \in \mathbb{N}\}$, which is a sequence satisfying

$$\hat{\mathfrak{X}}_l^j(m) \;=\; \begin{cases} \operatorname{sgn}(\Delta P_{jlm}) & \text{with probability } |\Delta P_{jlm}|, \\ 0 & \text{with probability } 1 - |\Delta P_{jlm}|, \end{cases}$$

for $m \in \mathbb{N}$. Letting $\bar{\mathfrak{X}}^j(m) = (\bar{\mathfrak{X}}_l^j(m))$ and $\hat{\mathfrak{X}}^j(m) = (\hat{\mathfrak{X}}_l^j(m))$, we have that

$$\mathfrak{X}^j(m) \;=\; \bar{\mathfrak{X}}^j(m) \;+\; \hat{\mathfrak{X}}^j(m) \tag{A.3}$$

for $j \in \mathcal{J}$ and $m \in \mathbb{N}$. Then, letting

$$\bar{\phi}^j(m) \;\coloneqq\; e^{i(j)} - \bar{\mathfrak{X}}^j(m) \quad \text{and} \quad \hat{\phi}^j(m) \;\coloneqq\; -\hat{\mathfrak{X}}^j(m), \tag{A.4}$$

we define

$$\bar{\Phi}^j(m) \;\coloneqq\; \bar{\phi}^j(1) \;+\; \cdots \;+\; \bar{\phi}^j(m) \quad \text{and} \quad \hat{\Phi}^j(m) \;\coloneqq\; \hat{\phi}^j(1) \;+\; \cdots \;+\; \hat{\phi}^j(m) \tag{A.5}$$

for $j \in \mathcal{J}$ and $m \in \mathbb{N}$. It follows from (A.1)–(A.2) and (A.3)–(A.5) that

$$\Phi^j(m) \;=\; \bar{\Phi}^j(m) \;+\; \hat{\Phi}^j(m) \tag{A.6}$$

for $j \in \mathcal{J}$ and $m \in \mathbb{N}$. We interpret $\bar{\Phi}^j = \{\bar{\Phi}^j(m) : m \in \mathbb{N}\}$ for $j \in \mathcal{J}$ as the cumulative routing process associated with activity j in a hypothetical system governed by transition probabilities $P_{jl}(\mu_j^\star, 0)$, i.e., those arising under the nominal service rate policy and no waiting times. Then, $\hat{\Phi}^j = \{\hat{\Phi}^j(m) : m \in \mathbb{N}\}$ for $j \in \mathcal{J}$ can be interpreted as the change in cumulative routing induced by dynamic service rate adjustments and system congestion.

Appendix B

Formal Derivation of the Approximating Brownian System

This section presents a formal derivation for the approximating Brownian control problem discussed in Chapter 4. The derivation presented here does not constitute a rigorous convergence proof of the pre-limit system to the approximating Brownian system. However, the arguments made in support of the Brownian approximation can be viewed as a broad outline for such a proof.

There are two common approaches to formally deriving approximating Brownian systems: weak convergence arguments and functional strong approximations. The former involves arguing that the distribution of a sequence of centered and scaled stochastic processes converges to the distribution of a set of limiting processes; see, e.g., Harrison (1988, 2003). For our problem, this approach would rely on the FCLT for Martingales; see, e.g., Whitt (2007, Theorem 2.1) and Ethier and Kurtz (2009, Chapter 7). Conversely, the latter involves arguing the almost sure convergence of a sequence of stochastic processes to a set of limiting processes; see, e.g., Çelik and Maglaras (2008) and Ata and Tongarlak (2013). While both methods ultimately produce the same approximating Brownian system, we adopt the functional strong approximations framework to simplify the exposition.

The general procedure is as follows: First, we consider a sequence of systems indexed by r, operating under heavy traffic conditions. We attach a superscript r to various quantities of interest corresponding to the r-th system. Second, we use strong approximation results to rewrite the various processes of interest. Finally, we let r get large and obtain the formal limit. We begin with the primitive processes.

B.1 Arrival, Abandonment, Service, Routing, and Flow Processes.

Arrival Processes. We begin with the arrival processes. For $i \in \mathcal{I}$, the class i (external) arrival process in the r-th system is a Poisson process with rate $r\lambda_i - \sqrt{r}\zeta_i$. Therefore, by the functional strong approximation for Poisson processes (see, e.g., Ethier and Kurtz (2009, Section 7.5)), we have that for $i \in \mathcal{I}$ and $t \geq 0$,

$$E_i^r(t) = r\lambda_i t - \sqrt{r}\zeta_i t + \sqrt{r}\hat{\chi}_i(t) + o(\sqrt{r}) \tag{B.1}$$

almost surely, where $\hat{\chi}_i = \{\hat{\chi}_i(t) : t \geq 0\}$ is a Brownian motion with zero drift and infinitesimal variance λ_i, and the notation $f(r) = o(g(r))$ implies that $f(r)/g(r) \to 0$ as $r \to \infty$. Let $\hat{\chi}(t) = (\hat{\chi}_i(t))$ for $t \geq 0$. It then follows from the independence of the arrival processes that $\hat{\chi} = \{\hat{\chi}(t) : t \geq 0\}$ is a Brownian motion with zero drift and covariance matrix $\mathrm{diag}(\lambda_1, \ldots, \lambda_I)$.

Abandonment Processes. The number of class i customers who have abandoned up to time t in the r-th system is given by

$$\Gamma_i^r(t) = N_i\Big(\int_0^t \ell_i^r Q_i^r(s)\, ds\Big),$$

where N_i is a unit rate Poisson process for $i \in \mathcal{I}$. It follows from (4.12), the functional strong approximation for Poisson processes, and the fact that $Q_i^r(t)$ is of order $\sqrt{r}$ that for $i \in \mathcal{I}$ and $t \geq 0$,

$$\Gamma_i^r(t) = \int_0^t \ell_i Q_i^r(s)\, ds + o(\sqrt{r}) \tag{B.2}$$

almost surely.

Service Processes. Next, we consider the service processes. It follows from (3.1) that the number of service completions of activity j up to time t in the r-th system is given by

$$S_j^r(t) = M_j\Big(\int_0^t n_j^r(s)\mu_j^r(s)ds\Big), \tag{B.3}$$

where M_j is a unit rate Poisson process for $j \in \mathcal{J}$. Substituting (4.5) and (4.7) into (B.3) yields

$$\begin{aligned} S_j^r(t) &= M_j\Big(\int_0^t n_j^\star\big(r\mu_j^\star + \sqrt{r}\eta_j(s)\big)ds - \frac{1}{\sqrt{r}}\int_0^t \big(r\mu_j^\star + \sqrt{r}\eta_j(s)\big)dY_j^r(s)\Big) \\ &= M_j\Big(rn_j^\star\mu_j^\star t + \sqrt{r}\int_0^t n_j^\star\eta_j(s)ds - \sqrt{r}\int_0^t \mu_j^\star dY_j^r(s) - \int_0^t \eta_j(s)dY_j^r(s)\Big) \end{aligned}$$

for $j \in \mathcal{J}$ and $t \geq 0$. It then follows from the functional strong approximation for Poisson processes that for $j \in \mathcal{J}$ and $t \geq 0$,

$$S_j^r(t) = rn_j^\star \mu_j^\star t + \sqrt{r}\int_0^t n_j^\star \eta_j(s)ds - \sqrt{r}\int_0^t \mu_j^\star dY_j^r(s) + \sqrt{r}\tilde{\chi}_j\Big(n_j^\star\mu_j^\star t + \frac{1}{\sqrt{r}}\int_0^t n_j^\star\eta_j(s)ds - \frac{1}{\sqrt{r}}\int_0^t \mu_j^\star dY_j^r(s) - \frac{1}{r}\int_0^t \eta_j(s)dY_j^r(s)\Big) + o(\sqrt{r}) \tag{B.4}$$

almost surely, where $\tilde{\chi}_j = \{\tilde{\chi}_j(t) : t \geq 0\}$ is a standard Brownian motion. Then, an (informal) application of the random time change theorem (on the fourth term) yields

$$S_j^r(t) = rn_j^\star\mu_j^\star t + \sqrt{r}\int_0^t n_j^\star\eta_j(s)ds - \sqrt{r}\int_0^t \mu_j^\star dY_j^r(s) + \sqrt{r}\tilde{\chi}_j(n_j^\star\mu_j^\star t) + o(\sqrt{r}). \tag{B.5}$$

Routing Processes. As a preliminary to the flow processes, we next rewrite the cumulative routing processes. In Appendix A.2, we show that the cumulative routing process associated with activity j, Φ^j, can be written as the sum of two independent processes $\bar{\Phi}^j$ and $\hat{\Phi}^j$; see (A.5) for their definitions. We interpret $\bar{\Phi}^j$ for $j \in \mathcal{J}$ as the cumulative routing process associated with activity j in a hypothetical system governed by transition probabilities $P_{jl}(\mu_j^\star, 0)$, i.e., those arising under the nominal service rate policy and no waiting times. We interpret $\hat{\Phi}^j$ for $j \in \mathcal{J}$ as the change in cumulative routing induced by dynamic service rate adjustments and system congestion. Since $\bar{\Phi}^j$ is the (partial) sum of independent and identically distributed random vectors, it follows from the multidimensional invariance principle, also known as strong approximation (see, e.g., Einmahl (1987a,b, 1989)), that for $j \in \mathcal{J}$ and $t \geq 0$,

$$\bar{\Phi}^{j,r}(rt) = \mathbb{E}[\bar{\Phi}^{j,r}(rt)] + \sqrt{r}\bar{\chi}^j(t) + o(\sqrt{r}) \tag{B.6}$$

almost surely, where $\bar{\chi}^j$ is an I-dimensional Brownian motion with zero drift and covariance matrix $\Omega^j = [\Omega_{il}^j]$ given by

$$\Omega_{il}^j := P_{ji}(\mu_j^\star, 0)\big(\mathbf{1}_{\{i=l\}} - P_{jl}(\mu_j^\star, 0)\big) \tag{B.7}$$

for $i, l \in \mathcal{I}$ and $j \in \mathcal{J}$. By (A.4), we have that $\hat{\phi}^j(m) = -\hat{\mathfrak{X}}^j(m)$, where $\hat{\mathfrak{X}}_l^j(m)$ is a Bernoulli

random variable with probability of success

$$\begin{aligned}\Delta P^r_{jlm} &= P^r_{jl}\big(\mu^r_j(t^{j,r}_m + w^r_{i(j)}(t^{j,r}_m)), w^r_{i(j)}(t^{j,r}_m)\big) - P_{jl}(\mu^\star_j, 0) \\ &= P_{jl}\big(\mu^\star_j + \frac{\eta_j(t^{j,r}_m + w^r_{i(j)}(t^{j,r}_m))}{\sqrt{r}}, w^r_{i(j)}(t^j_m)\big) - P_{jl}(\mu^\star_j, 0) \qquad \text{(B.8)} \\ &= \frac{1}{\sqrt{r}}\frac{\partial P_{jl}}{\partial \mu_j}(\mu^\star_j, 0)\eta_j(t^{j,r}_m + w^r_{i(j)}(t^j_m)) + \frac{\partial P_{jl}}{\partial w_{i(j)}}(\mu^\star_j, 0) w^r_{i(j)}(t^{j,r}_m) + o(\sqrt{r}), \qquad \text{(B.9)}\end{aligned}$$

where (B.8) follows from (4.3) and (B.9) follows from the multivariate version of Taylor's theorem. It follows from (3.8), (4.5), and (4.11) that η is bounded. Similarly, it follows from (3.9) and (4.11) that $\sqrt{r}w^r$ is bounded. Therefore, one can deduce from (B.9) that ΔP^r_{jlm} is of order $1/\sqrt{r}$. Then, it follows from Kolmogorov's Strong Law of Large Numbers (see, e.g., Feller (1991, Section 10.7)) that for $j \in \mathcal{J}$ and $t \geq 0$,

$$\hat{\Phi}^{j,r}(rt) = \mathbb{E}[\hat{\Phi}^{j,r}(rt)] + o(\sqrt{r}) \qquad \text{(B.10)}$$

almost surely. Therefore, by (A.6), (B.6), and (B.10), we have that for $j \in \mathcal{J}$ and $t \geq 0$,

$$\Phi^{j,r}(rt) = \mathbb{E}[\Phi^{j,r}(rt)] + \sqrt{r}\bar{\chi}^j(t) + o(\sqrt{r}) \qquad \text{(B.11)}$$

almost surely.

Flow Processes. We now rewrite the flow processes using the strong approximation results for the service and cumulative routing processes. For notational convenience, for $j \in \mathcal{J}$, we define the mapping $P^j : \mathcal{M}_j \times \mathbb{R}_+ \to [0,1]^I$ as follows: For $\mu_j \in \mathcal{M}_j, w_{i(j)} \in \mathbb{R}_+$,

$$P^j(\mu_j, w_{i(j)}) := \big(P_{j1}(\mu_j, w_{i(j)}), \ldots, P_{jI}(\mu_j, w_{i(j)})\big)'.$$

Here, $P^j(\mu_j, w_{i(j)})$ is the I-dimensional vector of transition probabilities associated with activity j at service rate μ_j and waiting time $w_{i(j)}$. It follows from (3.2) that for $j \in \mathcal{J}$ and $t \geq 0$,

$$\begin{aligned}F^{j,r}(t) &= \Phi^{j,r}(S^r_j(t)) = \sqrt{r}\bar{\chi}^j(S^r_j(t)/r) + \sum_{l=1}^{S^r_j(t)} \mathbb{E}[\phi^{j,r}(l)] + o(\sqrt{r}) \qquad \text{(B.12)} \\ &= \sqrt{r}\bar{\chi}^j\Big(n^\star_j\mu^\star_j t + \frac{1}{\sqrt{r}}\tilde{\chi}_j(n^\star_j\mu^\star_j t) + \frac{1}{\sqrt{r}}\int_0^t n^\star_j\eta_j(s)ds - \frac{1}{\sqrt{r}}\int_0^t \mu^\star_j dY^r_j(s) + o(\frac{1}{\sqrt{r}})\Big) \qquad \text{(B.13)} \\ &\quad + \int_0^t \big(e_{i(j)} - P^{j,r}(\mu^r_j(s), w^r_{i(j)}(s))\big)dS^r_j(s) + o(\sqrt{r}) \\ &= \sqrt{r}\bar{\chi}^j(n^\star_j\mu^\star_j t) + e_{i(j)}S^r_j(t) - \int_0^t P^{j,r}(\mu^r_j(s), w^r_{i(j)}(s))dS^r_j(s) + o(\sqrt{r}), \qquad \text{(B.14)}\end{aligned}$$

where (B.12) follows from (A.2) and (B.11), (B.13) follows from (A.1) and (B.5), and (B.14) follows from an (informal) application of the random time change theorem on the first term of (B.13). To further simplify (B.14), we next rewrite the third term on its right-hand side. It follows from (B.5) that

$$\begin{aligned}\int_0^t P^{j,r}(\mu_j^r(s), w_{i(j)}^r(s))dS_j^r(s) \;&= r\int_0^t P^{j,r}\big(\mu_j^r(s), w_{i(j)}^r(s)\big)n_j^\star\mu_j^\star ds + \sqrt{r}\int_0^t P^{j,r}\big(\mu_j^r(s), w_{i(j)}^r(s)\big)n_j^\star\eta_j(s)ds \\ &\quad - \sqrt{r}\int_0^t P^{j,r}\big(\mu_j^r(s), w_{i(j)}^r(s)\big)\mu_j^\star dY_j^r(s) \\ &\quad + \sqrt{r}\int_0^t P^{j,r}\big(\mu_j^r(s), w_{i(j)}^r(s)\big)d\tilde{\chi}_j(n_j^\star\mu_j^\star t) + o(\sqrt{r}) \qquad \text{(B.15)}\end{aligned}$$

for $j \in \mathcal{J}$ and $t \geq 0$. This expression can be simplified further. Specifically, it follows from (4.3), (4.5), and the multivariate version of Taylor's theorem that for $j \in \mathcal{J}$ and $t \geq 0$,

$$\begin{aligned}P^{j,r}\big(\mu_j^r(t), w_{i(j)}^r(t)\big) \;&= P^{j,r}\big(r\mu_j^\star + \sqrt{r}\eta_j(t), w_{i(j)}^r(t)\big) \\ &= P^j\big(\mu_j^\star + \frac{\eta_j(t)}{\sqrt{r}}, w_{i(j)}^r(t)\big) \\ &= P^j(\mu_j^\star, 0) + \frac{1}{\sqrt{r}}\nabla_\mu P^j(\mu_j^\star, 0)\eta_j(t) + \nabla_w P^j(\mu_j^\star, 0)w_{i(j)}^r(t) + o(\frac{1}{\sqrt{r}}),\end{aligned} \qquad \text{(B.16)}$$

where

$$\begin{aligned}\nabla_\mu P^j(\mu_j^\star, 0) \;&= \big(\partial P_{j1}(\mu_j^\star, 0)/\partial\mu_j, \ldots, \partial P_{jI}(\mu_j^\star, 0)/\partial\mu_j\big)', \\ \nabla_w P^j(\mu_j^\star, 0) \;&= \big(\partial P_{j1}(\mu_j^\star, 0)/\partial w_{i(j)}, \ldots, \partial P_{jI}(\mu_j^\star, 0)/\partial w_{i(j)}\big)'\end{aligned}$$

denote the Jacobians of P^j with respect to the service rate μ_j and the waiting time $w_{i(j)}$, respectively, evaluated at $(\mu_j^\star, 0)$ for $j \in \mathcal{J}$. Substituting (B.16) into (B.15) and simplifying the terms yields

$$\begin{aligned}\int_0^t P^{j,r}\big(\mu_j^r(s), w_{i(j)}^r(s)\big)dS_j^r(s) \;&= r\int_0^t P^j(\mu_j^\star, 0)n_j^\star\mu_j^\star ds + \sqrt{r}\int_0^t \nabla_\mu P^j(\mu_j^\star, 0)\eta_j(s)n_j^\star\mu_j^\star ds \\ &\quad + r\int_0^t \nabla_w P^j(\mu_j^\star, 0)w_{i(j)}^r(s)n_j^\star\mu_j^\star ds + \sqrt{r}\int_0^t P^j(\mu_j^\star, 0)n_j^\star\eta_j(s)ds \\ &\quad - \sqrt{r}P^j(\mu_j^\star, 0)\mu_j^\star Y_j^r(s) + \sqrt{r}P^j(\mu_j^\star, 0)\tilde{\chi}_j(n_j^\star\mu_j^\star t) + o(\sqrt{r})\end{aligned} \qquad \text{(B.17)}$$

for $j \in \mathcal{J}$ and $t \geq 0$. Then, substituting (B.5) and (B.17) into (B.14) yields

$$\begin{aligned} F^{j,r}(t) \;=\; & \sqrt{r}\bar{\chi}^{j}(n_j^\star \mu_j^\star t) \;-\; \sqrt{r}P^{j}(\mu_j^\star,0)\tilde{\chi}_j(n_j^\star \mu_j^\star t) \;-\; r\int_0^t P^{j}(\mu_j^\star,0) n_j^\star \mu_j^\star ds \\ & -\; \sqrt{r}\int_0^t \nabla_\mu P^{j}(\mu_j^\star,0)\eta_j(s) n_j^\star \mu_j^\star ds \;-\; r\int_0^t \nabla_w P^{j}(\mu_j^\star,0) w^r_{i(j)}(s) n_j^\star \mu_j^\star ds \\ & -\; \sqrt{r}\int_0^t P^{j}(\mu_j^\star,0) n_j^\star \eta_j(s) ds \;+\; \sqrt{r}P^{j}(\mu_j^\star,0)\mu_j^\star Y_j^r(t) \;+\; e_{i(j)} r n_j^\star \mu_j^\star t \\ & +\; e_{i(j)}\sqrt{r}\int_0^t n_j^\star \eta_j(s) ds \;-\; e_{i(j)}\sqrt{r}\mu_j^\star Y_j^r(t) \\ & +\; e_{i(j)}\sqrt{r}\tilde{\chi}_j(n_j^\star \mu_j^\star t) \;+\; o(\sqrt{r}) \end{aligned} \tag{B.18}$$

for $j \in \mathcal{J}$ and $t \geq 0$. Observe that the fifth term on the right-hand side of (B.18) contains $w^r_{i(j)}(t)$, the waiting time of a class $i(j)$ customer arriving at time t. Applying the "snapshot principle" of Reiman (1984), we can approximate $w_i^r(t)$ for $i \in \mathcal{I}$ as follows:

$$\sqrt{r}w_i^r(t) \;\approx\; \frac{Q_i^r(t)}{\sqrt{r}\sum_{j\in\mathcal{J}_i} n_j^\star \mu_j^\star} \;=\; \frac{Z_i^r(t)}{\sum_{j\in\mathcal{J}_i} n_j^\star \mu_j^\star}, \tag{B.19}$$

which becomes more accurate as the system approaches heavy traffic. The snapshot principle states that under heavy traffic conditions, the state of the system changes negligibly during a customer's sojourn in the system. In other words, one can take a "snapshot" of the system when a customer arrives, and queue lengths and waiting times of the various buffers change negligibly during the customer's sojourn in the system; see, Harrison and Nguyen (1990, 1993), Huang et al. (2015), and Kim et al. (2018) for similar treatments. Rearranging the terms in (B.18) and using (4.1) and (B.19) yields

$$\begin{aligned} F^{j,r}(t) \;=\; & \sqrt{r}\bar{\chi}^{j}(n_j^\star \mu_j^\star t) \;+\; \sqrt{r}(e_{i(j)} - P^{j}(\mu_j^\star,0))\tilde{\chi}_j(n_j^\star \mu_j^\star t) \;-\; r\int_0^t R^{j}(\mu^\star,0) n_j^\star ds \\ & -\; \sqrt{r}\int_0^t \nabla_\mu P^{j}(\mu_j^\star,0)\eta_j(s) n_j^\star \mu_j^\star ds \;-\; \sqrt{r}\int_0^t \frac{\nabla_w P^{j}(\mu_j^\star,0) Z^r_{i(j)}(s) n_j^\star \mu_j^\star}{\sum_{j'\in\mathcal{J}_{i(j)}} n_{j'}^\star \mu_{j'}^\star} ds \\ & +\; \sqrt{r}\int_0^t (e_{i(j)} - P^{j}(\mu_j^\star,0)) n_j^\star \eta_j(s) ds \;-\; \sqrt{r}R^{j}(\mu^\star,0) Y_j^r(t) \;+\; o(\sqrt{r}) \end{aligned} \tag{B.20}$$

for $j \in \mathcal{J}$ and $t \geq 0$.

B.2 Scaled Queue Length Process.

We next rewrite the scaled queue length process using the rewritten arrival and flow processes. For $t \geq 0$, we have that

$$\begin{aligned} Q^r(t) &= E^r(t) - \sum_{j\in\mathcal{J}} F^{j,r}(t) - \Gamma^r(t) - O^r(t) \\ &= -\sqrt{r}\zeta t + \sqrt{r}\hat{\chi}(t) - \sum_{j\in\mathcal{J}} \sqrt{r}\bar{\chi}^j(n_j^\star\mu_j^\star t) - \sum_{j\in\mathcal{J}} \sqrt{r}(e_{i(j)} - P^j(\mu_j^\star,0))\tilde{\chi}_j(n_j^\star\mu_j^\star t) \\ &\quad + \sqrt{r}\sum_{j\in\mathcal{J}} \int_0^t \nabla_\mu P^j(\mu_j^\star,0)\eta_j(s)n_j^\star\mu_j^\star ds + \sqrt{r}\sum_{j\in\mathcal{J}} \int_0^t \frac{\nabla_w P^j(\mu_j^\star,0)Z^r_{i(j)}(s)n_j^\star\mu_j^\star}{\sum_{j'\in\mathcal{J}_{i(j)}} n_{j'}^\star\mu_{j'}^\star} ds \\ &\quad - \sqrt{r}\sum_{j\in\mathcal{J}} \int_0^t (e_{i(j)} - P^j(\mu_j^\star,0))n_j^\star\eta_j(s)ds + \sqrt{r}\sum_{j\in\mathcal{J}} R^j(\mu^\star,0)Y_j^r(t) \\ &\quad - \int_0^t \ell' Q^r(s)\, ds - \sqrt{r}G^r(t) + o(\sqrt{r}), \end{aligned} \tag{B.21}$$

where (B.21) follows from (4.9), (B.1), (B.2), (B.20), and the fact that $(\mu^\star, n^\star)$ is an optimal solution to the SPP (4.2). Then, dividing both sides of (B.21) by $\sqrt{r}$ and using (4.9) yields

$$\begin{aligned} Z^r(t) &= -\zeta t + \hat{\chi}(t) - \sum_{j\in\mathcal{J}} \bar{\chi}^j(n_j^\star\mu_j^\star t) - \sum_{j\in\mathcal{J}} (e_{i(j)} - P^j(\mu_j^\star,0))\tilde{\chi}_j(n_j^\star\mu_j^\star t) \\ &\quad + \sum_{j\in\mathcal{J}} \int_0^t \nabla_\mu P^j(\mu_j^\star,0)\eta_j(s)n_j^\star\mu_j^\star ds + \sum_{j\in\mathcal{J}} \int_0^t \frac{\nabla_w P^j(\mu_j^\star,0)Z^r_{i(j)}(s)n_j^\star\mu_j^\star}{\sum_{j'\in\mathcal{J}_{i(j)}} n_{j'}^\star\mu_{j'}^\star} ds \\ &\quad - \sum_{j\in\mathcal{J}} \int_0^t (e_{i(j)} - P^j(\mu_j^\star,0))n_j^\star\eta_j(s)ds + \sum_{j\in\mathcal{J}} R^j(\mu^\star,0)Y_j^r(t) \\ &\quad - \int_0^t \ell' Z^r(s)ds - G^r(t) + o(1) \end{aligned} \tag{B.22}$$

for $t \geq 0$. We define the process $X = \{X(t) : t \geq 0\}$ as follows: For $t \geq 0$,

$$X(t) = -\zeta t + \hat{\chi}(t) - \sum_{j\in\mathcal{J}} \bar{\chi}^j(n_j^\star\mu_j^\star t) - \sum_{j\in\mathcal{J}} (e_{i(j)} - P^j(\mu_j^\star,0))\tilde{\chi}_j(n_j^\star\mu_j^\star t). \tag{B.23}$$

Since $\hat{\chi}$, $\bar{\chi}^j$, and $\tilde{\chi}_j$ for $j \in \mathcal{J}$ are independent Brownian motions with zero drift, X is a Brownian motion with drift $-\zeta$ and covariance matrix

$$\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_I) + \sum_{j\in\mathcal{J}} \left(\mu_j^\star n_j^\star \Omega^j + \frac{n_j^\star}{\mu_j^\star} R^j(\mu^\star,0) R^j(\mu^\star,0)'\right),$$

where R is given by (4.1). To simplify the analysis to follow, we define the $I \times J$ matrix $C = [C_{ij}]$ and the $I \times I$ matrix $D = [D_{il}]$ as follows: $i, l \in \mathcal{I}$ and $j \in \mathcal{J}$,

$$C_{ij} \coloneqq \left(\mathbf{1}_{\{i(j)=i\}} - P_{ji}(\mu_j^\star, 0)\right) n_j^\star - n_j^\star \mu_j^\star \frac{\partial P_{ji}}{\partial \mu_j}(\mu_j^\star, 0), \quad D_{il} \coloneqq \ell_i \mathbf{1}_{\{i=l\}} - \sum_{j \in \mathcal{J}_l} \frac{n_j^\star \mu_j^\star}{\sum_{j' \in \mathcal{J}_l} n_{j'}^\star \mu_{j'}^\star} \frac{\partial P_{ji}}{\partial w_{i(j)}}(\mu_j^\star, 0). \tag{B.24}$$

Then, substituting (B.23)–(B.24) into (B.22) yields

$$Z^r(t) = X(t) - \int_0^t C\eta(s)ds - \int_0^t DZ^r(s)ds + R(\mu^\star, 0)Y^r(t) - G^r(t) + o(1) \tag{B.25}$$

for $t \geq 0$.

B.3 Scaled Age Process.

We next rewrite the scaled age process. Since the waiting times are of order $1/\sqrt{r}$ and the service times are of order $1/r$ (see (4.5)), it follows from the "snapshot principle" of Reiman (1984) that for $i \in \mathcal{I}$, $l \in \mathcal{B}_i$, and $t \geq 0$

$$a_l^r(t) = a_i^r(t) + w_l^r(t) + o(1/\sqrt{r}). \tag{B.26}$$

To reiterate, the snapshot principle states that under heavy traffic conditions, the state of the system changes negligibly during a customer's sojourn in the system. In other words, one can take a "snapshot" of the system when a customer arrives, and queue lengths and waiting times of the various buffers change negligibly during the customer's sojourn in the system; see, Harrison and Nguyen (1990, 1993), Huang et al. (2015), and Kim et al. (2018) for similar treatments. Multiplying both sides of (B.26) by $\sqrt{r}$ and using (4.10) and (B.19) yields

$$\alpha_l^r(t) = \alpha_i^r(t) + \frac{Z_l^r(t)}{\sum_{j \in \mathcal{J}_l} n_j^\star \mu_j^\star} + o(1) \tag{B.27}$$

for $i \in \mathcal{I}$, $l \in \mathcal{B}_i$, and $t \geq 0$. We can rewrite (B.27) in vector form as

$$\alpha^r(t) = \Pi\alpha^r(t) + \operatorname{diag}\left(\frac{1}{\sum_{j \in \mathcal{J}_1} n_j^\star \mu_j^\star}, \frac{1}{\sum_{j \in \mathcal{J}_2} n_j^\star \mu_j^\star}, \ldots, \frac{1}{\sum_{j \in \mathcal{J}_I} n_j^\star \mu_j^\star}\right) Z^r(t) + o(1)$$

for $t \geq 0$, where the $I \times I$ matrix $\Pi = [\Pi_{il}]$ is defined as follows: For $i, l \in \mathcal{I}$,

$$\Pi_{il} \coloneqq \begin{cases} 1, & \text{if } i \in \mathcal{B}_l, \\ 0, & \text{otherwise.} \end{cases}$$

Then, rearranging the terms yields

$$\alpha^r(t) \;=\; \Lambda Z^r(t) \;+\; o(1) \tag{B.28}$$

for $t \geq 0$, where the nonnegative $I \times I$ matrix Λ is given by

$$\Lambda \;\coloneqq\; (I - \Pi)^{-1} \operatorname{diag}\big(\frac{1}{\sum_{j \in \mathcal{J}_1} n_j^\star \mu_j^\star}, \frac{1}{\sum_{j \in \mathcal{J}_2} n_j^\star \mu_j^\star}, \ldots, \frac{1}{\sum_{j \in \mathcal{J}_I} n_j^\star \mu_j^\star}\big).$$

B.4 Centered and Scaled Cumulative Cost Process.

We next rewrite the centered and scaled cumulative cost process. For $t \geq 0$, we have that

$$\begin{aligned} V^r(t) \;&=\; \sum_{j \in \mathcal{J}} \int_0^t n_j^r(s) \mu_j^r(s)\, \upsilon_j^r\big(\mu_j^r(s), w_{i(j)}^r(s)\big) ds \;+\; \sum_{i \in \mathcal{I}} \pi_i^r \Gamma_i^r(t) \;+\; \sum_{i \in \mathcal{I}} \nu_i^r O_i^r(t) \\ &=\; \sum_{j \in \mathcal{J}} \int_0^t n_j^\star\big(r\mu_j^\star + \sqrt{r}\eta_j(s)\big) \upsilon_j\big(\mu_j^\star + \frac{\eta_j(s)}{\sqrt{r}}, w_{i(j)}^r(s)\big) ds \\ &\quad - \sum_{j \in \mathcal{J}} \frac{1}{\sqrt{r}} \int_0^t \big(r\mu_j^\star \;+\; \sqrt{r}\eta_j(s)\big) \upsilon_j\big(\mu_j^\star + \frac{\eta_j(s)}{\sqrt{r}}, w_{i(j)}^r(s)\big) dY_j^r(s) \\ &\quad + \sum_{i \in \mathcal{I}} \pi_i^r \Gamma_i^r(t) \;+\; \sum_{i \in \mathcal{I}} \nu_i^r O_i^r(t), \end{aligned} \tag{B.29}$$

where (B.29) follows from (4.3)–(4.7). To further simplify (B.29), we apply the multivariate version of Taylor's theorem, which yields, for $j \in \mathcal{J}$ and $t \geq 0$,

$$\upsilon_j\big(\mu_j^\star + \frac{\eta_j(t)}{\sqrt{r}}, w_{i(j)}^r(t)\big) \;=\; \upsilon_j(\mu_j^\star, 0) \;+\; \frac{\partial \upsilon_j}{\partial \mu_j}(\mu_j^\star, 0) \frac{\eta_j(t)}{\sqrt{r}} \;+\; \frac{\partial \upsilon_j}{\partial w_{i(j)}}(\mu_j^\star, 0) w_{i(j)}^r(t) \;+\; o(1/\sqrt{r}). \tag{B.30}$$

Substituting (4.5) and (B.30) into (B.29), rearranging the terms, and using (4.3), (4.9), (4.12), and (B.19) yields

$$\begin{aligned} V^r(t) \;=\; & r \sum_{j\in\mathcal{J}} n_j^\star \mu_j^\star \upsilon_j(\mu_j^\star, 0) t \;+\; \sqrt{r} \sum_{j\in\mathcal{J}} \int_0^t n_j^\star \mu_j^\star \frac{\partial \upsilon_j}{\partial \mu_j}(\mu_j^\star, 0) \eta_j(s) ds \\ & + \sqrt{r} \sum_{j\in\mathcal{J}} \int_0^t n_j^\star \mu_j^\star \frac{\partial \upsilon_j}{\partial w_{i(j)}}(\mu_j^\star, 0) \frac{Z^r_{i(j)}(s)}{\sum_{j'\in\mathcal{J}_{i(j)}} n_{j'}^\star \mu_{j'}^\star} ds \\ & + \sqrt{r} \sum_{j\in\mathcal{J}} \int_0^t n_j^\star \eta_j(s) \upsilon_j(\mu_j^\star, 0) ds \;-\; \sqrt{r} \sum_{j\in\mathcal{J}} \mu_j^\star \upsilon_j(\mu_j^\star, 0) Y_j^r(t) \\ & + \sqrt{r} \sum_{i\in\mathcal{I}} \int_0^t \pi_i \ell_i Z_i^r(s)\, ds \;+\; \sqrt{r} \sum_{i\in\mathcal{I}} \nu_i G_i^r(t) \;+\; o(\sqrt{r}). \end{aligned} \tag{B.31}$$

For ease of exposition, we define the I-dimensional vector $b = (b_j)$ and I-dimensional vector $\mathfrak{g} = (\mathfrak{g}_i)$ as follows: For $i \in \mathcal{I}$ and $j \in \mathcal{J}$,

$$b_j \;\coloneqq\; n_j^\star \upsilon_j(\mu_j^\star, 0) \;+\; n_j^\star \mu_j^\star \frac{\partial \upsilon_j}{\partial \mu_j}(\mu_j^\star, 0) \quad \text{and} \quad \mathfrak{g}_i \;\coloneqq\; \pi_i \ell_i \;+\; \sum_{j\in\mathcal{J}_i} \frac{n_j^\star \mu_j^\star}{\sum_{j'\in\mathcal{J}_i} n_{j'}^\star \mu_{j'}^\star} \frac{\partial \upsilon_j}{\partial w_{i(j)}}(\mu_j^\star, 0). \tag{B.32}$$

Then, it follows from (B.31)–(B.32) that the centered and scaled cumulative cost process for the r-th system is given by

$$\xi^r(t) \;=\; \int_0^t b'\eta(s) ds \;+\; \int_0^t \mathfrak{g}' Z^r(s) ds \;-\; \sum_{j\in\mathcal{J}} \mu_j^\star \upsilon_j(\mu_j^\star, 0) Y_j^r(t) \;+\; \sum_{i\in\mathcal{I}} \nu_i G_i^r(t) \;+\; o(1). \tag{B.33}$$

B.5 Approximating Brownian System

We established that the scaled queue length process satisfies (B.25), the scaled age process satisfies (B.28), and the centered and scaled cumulative cost process satisfies (B.33). Next, we argue similar to Harrison (1988, 2000, 2003) that as the system gets large, the scaled processes defined above converge to their formal limits $\xi, Z, \mathfrak{U}, Y$, and G that jointly satisfy the following for $t \geq 0$, and some

$\mu_N^\star \in \mathcal{M}_N$ and $\Delta_0 \in \mathbb{R}^J$:

$$\xi(t) = \int_0^t b'\eta(s)ds + \int_0^t \mathfrak{g}'Z(s)ds - \sum_{j\in\mathcal{J}} \mu_j^\star \upsilon_j(\mu_j^\star, 0)Y_j(t) + \sum_{i\in\mathcal{I}} \nu_i G_i(t),$$

$$Z(t) = X(t) - \int_0^t C\eta(s)ds - \int_0^t DZ(s)ds + R(\mu^\star, 0)Y(t) - G(t),$$

$$\alpha(t) = \Lambda Z(t),$$

$$\mathfrak{U}(t) = HY(t),$$

$\mathfrak{U}$ and G are nondecreasing with $\mathfrak{U}(0) = G(0) = 0$,

$$\Delta_0 - \Delta/2 \le \eta(t) \le \Delta_0 + \Delta/2,$$

$\alpha(t) \le \delta$ and $Z(t) \ge 0$.

Appendix C

Proofs of the Technical Results

This section contains the proofs of the technical results. Section C.1 analyzes the IVP associated with the Bellman equation and establishes key properties of its solution. Section C.2 contains the proofs of the main technical results. Section C.3 compiles auxiliary lemmas used throughout.

C.1 An Initial Value Problem Related to the Bellman Equation

This section develops some analytical properties for the initial value problem (IVP) (6.16)–(6.17) introduced in Chapter 6 as part of the analysis of the workload formulation. For completeness, we restate the problem here: For $\gamma \in [\underline{\gamma}, \infty)$, find a function $v \in C^1[l_\gamma, \bar{w}]$ that satisfies

$$\frac{1}{2}\sigma^2 v'(w) + \mu v(w) + \psi(v(w)) + \phi(w, v(w)) = \gamma, \quad w \in [l_\gamma, \bar{w}], \tag{C.1}$$

subject to the boundary condition

$$v(l_\gamma) = -\kappa_L, \tag{C.2}$$

where l_γ is given by (6.15). The following lemma establishes the existence and uniqueness of a solution to the IVP (C.1)–(C.2).

Lemma C.1. *For each $\gamma \in \mathbb{R}$, the IVP (C.1)–(C.2) admits a unique solution $v_\gamma \in C^1([l_\gamma, \bar{w}])$.*

Proof. Fix $\gamma \in \mathbb{R}$ and define $f : [l_\gamma, \bar{w}] \times \mathbb{R} \to \mathbb{R}$ as follows:

$$f(w, y) \coloneqq \frac{2}{\sigma^2}\big(-\mu y - \psi(y) - \phi(w, y) + \gamma\big), \quad (w, y) \in [l_\gamma, \bar{w}] \times \mathbb{R}.$$

The IVP (C.1)–(C.2) can be written as $v'(w) = f(w, v(w))$ for $w \in [l_\gamma, \bar{w}]$ subject to the boundary condition $v(l_\gamma) = -\kappa_L$. It follows from Lemma 6.3 that f is locally Lipschitz in y. Thus, by the Picard–Lindelöf Theorem, the IVP (C.1)–(C.2) admits a unique solution $v_\gamma \in C^1([l_\gamma, \bar{w}])$. □ □

The next result develops some key structural properties for the function v_γ, which play a central role in establishing the existence and uniqueness of a solution to the Bellman equation.

Lemma C.2. *The mapping $(w, \gamma) \mapsto v_\gamma(w)$ is jointly continuous for $\gamma \in [\underline{\gamma}, \infty)$ and $w \in [l_\gamma, \bar{w}]$. Moreover, the mapping $\gamma \mapsto v_\gamma(w)$ is strictly increasing in the following sense: For $\gamma_1, \gamma_2 \in [\underline{\gamma}, \infty)$ such that $\gamma_1 < \gamma_2$, we have that $v_{\gamma_1}(w) < v_{\gamma_2}(w)$ for $w \in [l_{\gamma_1}, \bar{w}]$. Furthermore, $v_{\underline{\gamma}}(w) < \kappa_U$ for $w \in [\underline{w}, \bar{w}]$.*

Proof. The joint continuity of the mapping $(w, \gamma) \mapsto v_\gamma(w)$ follows from Coddington and Levinson (1955, Chapter 2, Theorem 4.3); see also Lemmas 6.3 and 6.5. To establish strict monotonicity, we rearrange the terms in (C.1)–(C.2) to obtain

$$\frac{1}{2}\sigma^2 v'(w) = \gamma - \mu v(w) - \psi(v(w)) - \phi(w, v(w)), \quad w \in [l_\gamma, \bar{w}] \tag{C.3}$$

subject to the boundary condition

$$v(l_\gamma) = -\kappa_L. \tag{C.4}$$

Consider $\gamma_1, \gamma_2 \in [\underline{\gamma}, \infty)$ with $\gamma_1 < \gamma_2$. By Lemma 6.5, we have $l_{\gamma_1} \geq l_{\gamma_2}$. We want to show that $v_{\gamma_1}(w) < v_{\gamma_2}(w)$ for $w \in [l_{\gamma_1}, \bar{w}]$. We now consider the two possibilities: (1) $l_{\gamma_1} = l_{\gamma_2}$ and (2) $l_{\gamma_1} > l_{\gamma_2}$.

Case 1. Suppose $l_{\gamma_1} = l_{\gamma_2}$. It follows from the definition of γ in (6.15) that $l_{\gamma_1} = l_{\gamma_2} = \underline{w}$. Substituting this into (C.4) yields $v(l_{\gamma_1}) = v(l_{\gamma_2}) = -\kappa_L$. Then, it follows from (C.3) and the fact that ψ, ϕ, v_{γ_1} and v_{γ_2} are continuous and $\gamma_1 < \gamma_2$ that there exists $\epsilon > 0$ such that $v'_{\gamma_1}(w) < v'_{\gamma_2}(w)$ for $w \in (\underline{w}, \underline{w} + \epsilon)$, and hence $v_{\gamma_1}(w) < v_{\gamma_2}(w)$ for $w \in (\underline{w}, \underline{w} + \epsilon)$. Applying Lemma C.3 (in Appendix C.3) at $w = \underline{w} + \epsilon$ yields $v_{\gamma_1}(w) < v_{\gamma_2}(w)$ for $w \in (\underline{w}, \bar{w}]$ as desired.

Case 2. Suppose $l_{\gamma_1} > l_{\gamma_2}$. Since $l_{\gamma_1} > \underline{w}$, it follows from (6.15) and (C.3) that

$$\gamma_1 = -\mu\kappa_L + \psi(-\kappa_L) + \phi(l_{\gamma_1}, -\kappa_L), \tag{C.5}$$

$v'_{\gamma_1}(l_{\gamma_1}) = 0$, and $w \mapsto \phi(w, -\kappa_L)$ is nonincreasing on $[\underline{w}, l_{\gamma_1})$. Then, since $w \mapsto \phi(w, -\kappa_L)$ is convex and $l_{\gamma_1} > l_{\gamma_2}$, we conclude that

$$\gamma_2 \geq -\mu\kappa_L + \psi(-\kappa_L) + \phi(l_{\gamma_2}, -\kappa_L). \tag{C.6}$$

It follows from the definition of l_{γ_1} in (6.15) that $w \mapsto \phi(w, -\kappa_L)$ is strictly decreasing on $[l_{\gamma_2}, l_{\gamma_1}]$. Therefore, by Lemma C.5 (in Appendix C.3), we have that $v_{\gamma_2}(l_{\gamma_1}) > v_{\gamma_1}(l_{\gamma_1}) = -\kappa_L$. Then, applying Lemma C.3 (in Appendix C.3) at $w = l_{\gamma_1}$ yields $v_{\gamma_2}(w) > v_{\gamma_1}(w)$ for $w \in (l_{\gamma_1}, \bar{w}]$ as desired.

It remains to show that $v_{\underline{\gamma}}(w) < \kappa_U$ for $w \in [l_{\underline{\gamma}}, \bar{w}]$. It follows from the definition of $\underline{\gamma}$ in (6.13) that $\underline{\gamma} = \mu\kappa_L + \psi(-\kappa_L) + \phi(l_{\underline{\gamma}}, -\kappa_L)$ and the mapping $w \mapsto \phi(w, -\kappa_L)$ is nondecreasing on $[l_{\underline{\gamma}}, \bar{w}]$. Let

$$w_0 \coloneqq \inf\left\{w : \phi(w, -\kappa_L) > \phi(l_{\underline{\gamma}}, -\kappa_L), w \in [l_{\underline{\gamma}}, \bar{w}]\right\}. \tag{C.7}$$

It follows from (C.7) and convexity of $w \mapsto \phi(w, -\kappa_L)$ that $\phi(w, -\kappa_L) = \phi(l_{\underline{\gamma}}, -\kappa_L)$ for $w \in [l_{\underline{\gamma}}, w_0]$ and $w \mapsto \phi(w, -\kappa_L)$ is strictly increasing on $(w_0, \bar{w}]$. Then, by (C.1)–(C.2), we have that $v_{\underline{\gamma}}(w) = -\kappa_L$ for $w \in [l_{\underline{\gamma}}, w_0]$, and by Lemma C.4 (in Appendix C.3), we have that $v_{\underline{\gamma}}(w) \leq -\kappa_L$ for $w \in [w_0, \bar{w}]$. Combining these yields $v_{\underline{\gamma}}(w) \leq -\kappa_L < \kappa_U$ for $w \in [l_{\underline{\gamma}}, \bar{w}]$ as desired. □

□

C.2 Miscellaneous Proofs

Proof of Proposition 5.3. The proof consists of two parts. Part 1 establishes the structural properties of m and g, and Part 2 establishes the structural properties of y and ρ.

Part 1. By Assumption 5.2, there are $J_b = I + K - 1$ basic activities and all servers communicate. Therefore, each server has at most two basic activities and each buffer is served by at most two basic activities. It then follows from the definition of H in (4.4) and R in (4.1) that the the matrix $[H'\ R(\mu^\star, 0)']$ has full rank. Therefore, the workload space

$$\mathbb{M} \coloneqq \{a \in \mathbb{R}^I : a'R(\mu^\star, 0) = b'H \text{ for some } b \in \mathbb{R}^{K+J-J_b}\}$$

is one-dimensional (since $a'R(\mu^\star, 0) = b'H$ is a system of linear equations with J equations and $K + J - J_b + I = J + 1$ unknowns and its coefficient matrix has full rank). Let m be a nonzero vector spanning this space. Since the rows of H are linearly independent, this relation uniquely determines g for any given $m \in \mathbb{M}$. Since the vector m spans a one-dimensional space, it is unique up to a multiplicative constant, which implies that the vector g is also unique up to a multiplicative constant.

Part 2. We now turn to the equation

$$-(\mu_j^\star \upsilon_j(\mu_j^\star, 0))' \;=\; \rho' H \;+\; y' R(\mu^\star, 0). \tag{C.8}$$

This is a system of linear equations with J equations and $K + J - J_b + I = J + 1$ unknowns, and its coefficient matrix has full rank (as established earlier). Therefore, there exists a pair (y, ρ) that satisfies (C.8), and the solution set is one-dimensional. To show that there exists a solution with nonnegative ρ, consider the following linear problem: Choose $n \in \mathbb{R}_+^J$ and $o \in \mathbb{R}_+^I$ so as to

$$\text{minimize} \;\sum_{j \in \mathcal{J}} n_j \mu_j^\star \upsilon_j(\mu_j^\star, 0) \;+\; \sum_{i \in \mathcal{I}} \nu_i o_i \;\; \text{subject to } R(\mu^\star, 0)\, n \;+\; o \;=\; \lambda \text{ and } An \;\leq\; e. \tag{C.9}$$

The dual of this linear program is given as follows: Choose $\alpha \in \mathbb{R}^I$ and $\beta \in \mathbb{R}_+^K$ so as to

$$\text{maximize} \;\; \alpha'\lambda \;-\; \beta' e \;\; \text{subject to } \alpha' R(\mu^\star, 0) \;\leq\; (\mu_j^\star \upsilon_j(\mu_j^\star, 0))' \;+\; \beta' A \text{ and } \alpha \leq \nu, \tag{C.10}$$

where $\nu = (\nu_i)$ is the vector of rejection costs. The vectors α and β are the dual variables associated with the flow balance and capacity constraints in (C.9), respectively. It follows from Assumption 4.1 that (C.9) is feasible and bounded and $(n^\star, o^\star)$ is its unique optimal solution. Therefore, (C.10) has an optimal solution, which we denote by (α, β). By complementary slackness, the dual constraint for basic activities is binding at the optimal dual solution, i.e.,

$$\alpha' R^j(\mu^\star, 0) \;=\; \mu_j^\star \upsilon_j(\mu_j^\star, 0) \;+\; (\beta' A)_j, \quad j = 1, \ldots, J_b, \tag{C.11}$$

and it is nonbinding for nonbasic activities at the optimal dual solution, i.e.,

$$\alpha' R^j(\mu^\star, 0) \;<\; \mu_j^\star \upsilon_j(\mu_j^\star, 0) \;+\; (\beta' A)_j, \quad j = J_b + 1, \ldots, J. \tag{C.12}$$

We extend β from $\mathbb{R}_+^K$ to $\mathbb{R}_+^{K+J-J_b}$ by appending to the original dual variable vector the slack values from the nonbasic constraints. Then, it follows from (4.4) and (C.11)–(C.12) that the extended vector β satisfies $-(\mu_j^\star \upsilon_j(\mu_j^\star, 0)) = -\alpha' R(\mu^\star, 0) + \beta' H$. Then, setting $\rho = \beta$ and $y = -\alpha$, we have constructed a solution to (C.8) with $\rho \geq 0$. It also follows from the second constraint in the dual problem that $y + \nu \geq 0$. □

Proof of Proposition 5.4. The proof consists of two parts. In Part 1, we construct an admissible policy for the EWF (5.17) from a given admissible policy for the BCP (4.24), and show that its long-run average cost is less than or equal to that of the BCP policy. In Part 2, we construct an admissible policy for the BCP (4.24) from a given admissible policy for the EWF (5.17), and show its long-run average cost is equal to that of the EWF policy.

Part 1. Let (η, Y, G) be an admissible policy for the BCP (4.24), with the corresponding state descriptor $Z = \{Z(t) : t \geq 0\}$ given by (4.18). We will show that there exists an admissible policy (L, U, z, θ) for the EWF (5.17) with a long-run average cost less than or equal to that of the policy (η, Y, G) for the BCP (4.24). To that end, we define

$$L(t) \coloneqq g^{+\prime} HY(t) + m^{-\prime} G(t), \quad U(t) \coloneqq g^{-\prime} HY(t) + m^{+\prime} G(t), \quad \theta(t) \coloneqq m' C\eta(t), \tag{C.13}$$

for $t \geq 0$. Moreover, we define $z(t, w) \coloneqq Z(t)$ for $t \geq 0$ and $w \in \mathbb{R}$ and $W(t) := m'Z(t)$ for $t \geq 0$. Then, it follows from (4.18) and (C.13) that

$$W(t) = B(t) - \int_0^t \theta(s)\, ds - \int_0^t d' z\big(s, W(s)\big)\, ds + L(t) - U(t), \quad t \geq 0. \tag{C.14}$$

By (4.19)–(4.23) and (C.14), it follows that (5.13)–(5.16) hold. Hence, (L, U, z, θ) is an admissible policy for EWF (5.17). Finally, by (5.7) and (5.11), we have that

$$c(\theta(t)) \leq \beta' \eta(t) \quad \text{and} \quad \kappa_L L(t) + \kappa_U U(t) \leq \rho' HY(t) + (y + \nu)' G(t), \quad t \geq 0. \tag{C.15}$$

Then, we have that for $t \geq 0$,

$$\begin{aligned}
\Xi(t) &= \int_0^t c(\theta(s))ds + \int_0^t h' z(s, W(s))ds + \kappa_L L(t) + \kappa_U U(t) && \text{(C.16)}\\
&\leq \int_0^t \beta'\eta(s)ds + \int_0^t h' Z(s)ds + \rho' HY(t) + (y+\nu)' G(t) && \text{(C.17)}\\
&= \int_0^t (b + C'y)'\eta(s)ds + \int_0^t (\mathfrak{g} + D'y)' Z(s)ds + \rho' HY(t) + (y+\nu)' G(t) && \text{(C.18)}\\
&= \int_0^t b'\eta(s)ds + \int_0^t \mathfrak{g}' Z(s)ds + \int_0^t y' C\eta(s)ds + \int_0^t y' DZ(s)ds + y' G(t) + \rho' HY(t) + \nu' G(t) && \text{(C.19)}\\
&= \int_0^t b'\eta(s)ds + \int_0^t \mathfrak{g}' Z(s)ds + y' X(t) - y' Z(t) + y' R(\mu^\star, 0) Y(t) + \rho' HY(t) + \nu' G(t) && \text{(C.20)}\\
&= \int_0^t b'\eta(s)ds + \int_0^t \mathfrak{g}' Z(s)ds + y' X(t) - y' Z(t) - \sum_{j \in \mathcal{J}} \mu_j^\star v_j(\mu_j^\star, 0) Y_j(t) + \sum_{i \in \mathcal{I}} \nu_i G_i(t). && \text{(C.21)}
\end{aligned}$$

where (C.16) follows from (5.12), (C.17) follows from (C.15), (C.18) follows from (5.6) and (5.10), (C.19) follows by rearranging the terms, (C.20) follows from (4.18), (C.21) follows from (5.1). Then, taking expectation of both sides of (C.21), dividing by t, and taking lim sup as $t \to \infty$ yields the

following: For $t \geq 0$,

$$\limsup_{t\to\infty} \frac{1}{t}\,\mathbb{E}\big[\Xi(t)\big] \;\leq\; \limsup_{t\to\infty} \frac{1}{t}\,\mathbb{E}\big[\xi(t)\big] \;+\; \limsup_{t\to\infty} \frac{1}{t}\,\mathbb{E}\big[y'X(t)\big] \;-\; \limsup_{t\to\infty} \frac{1}{t}\,\mathbb{E}\big[y'Z(t)\big] \tag{C.22}$$

$$= \limsup_{t\to\infty} \frac{1}{t}\,\mathbb{E}\big[\xi(t)\big] \;-\; y'\zeta, \tag{C.23}$$

where (C.22) follows from (4.17) and (C.23) follows from the fact that X is a Brownian motion with drift vector $-\zeta$, and Z is bounded.

Part 2. Conversely, let (L, U, z, θ) be an admissible policy for the EWF (5.17), with the corresponding state descriptor $W(t) = \{W(t) : t \geq 0\}$ given by (5.13). We will show that there exists an admissible policy (η, Y, G) for the BCP (4.24) with a long-run average cost equal to that of the policy (L, U, z, θ) for the EWF (5.17). To that end, given the process L and U, we define G and $\mathfrak{U}$ by "lifting" L and U in the obvious way: allocate all increments of L (respectively, U) to whichever rejection, idleness, or nonbasic activity that attains the minimum effective cost per unit increase (respectively, decrease) in the workload process; see (5.5) and (5.11). Moreover, we define

$$\eta(t) \;\coloneqq\; \eta^\star(\theta(t)), \quad t \geq 0 \tag{C.24}$$

where $\eta^\star$ is given by (5.7), and define the state descriptor $Z = \{Z(t) : t \geq 0\}$ as $Z(t) \coloneqq z(t, W(t))$ for $t \geq 0$. It remains to specify the process $Y = \{Y(t) : t \geq 0\}$. We would like Y to satisfy

$$R(\mu^\star, 0)Y(t) \;=\; Z(t) \;-\; X(t) \;+\; \int_0^t C\eta(s)ds \;+\; \int_0^t DZ(s)ds \;+\; G(t) \quad \text{and} \quad \mathfrak{U}(t) = HY(t), \quad t \geq 0. \tag{C.25}$$

It follows from Harrison and Williams (2005, Lemma 4.4) such a process Y exists. Then, it follows from (4.17)–(4.23) and (C.24)–(C.25) that (η, Y, G) is an admissible policy for BCP (4.24). Finally,

we have that for $t \geq 0$,

$$\xi(t) = \int_0^t b'\eta(s)ds + \int_0^t \mathfrak{g}'Z(s)ds - \sum_{j\in\mathcal{J}} \mu_j^\star v_j(\mu_j^\star, 0)Y_j(t) + \nu'G(t) \tag{C.26}$$

$$= \int_0^t b'\eta(s)ds + \int_0^t \mathfrak{g}'Z(s)ds + y'R(\mu^\star, 0)Y(t) + \rho'HY(t) + \nu'G(t) \tag{C.27}$$

$$= \int_0^t b'\eta(s)ds + \int_0^t \mathfrak{g}'Z(s)ds - y'X(t) + y'Z(t) + \int_0^t y'C\eta(s)ds + \int_0^t y'DZ(s)ds + y'G(t) + \rho'HY(t) + \nu'G(t) \tag{C.28}$$

$$= \int_0^t (b + C'y)'\eta(s)ds + \int_0^t (\mathfrak{g} + D'y)'Z(s)ds - y'X(t) + y'Z(t) + \rho'HY(t) + (y + \nu)'G(t) \tag{C.29}$$

$$= \int_0^t c(\theta(s))ds + \int_0^t h'z(s, W(s))ds - y'X(t) + y'z(t, W(t)) + \kappa_L L(t) + \kappa_U U(t) \tag{C.30}$$

where (C.26) follows from (4.17), (C.27) follows from (5.1), (C.28) follows from (4.18), (C.29) follows by rearranging the terms, (C.30) follows from (5.6), (5.7), and (5.10). Then, taking expectation of both sides of (C.30), dividing by t, and taking lim sup as $t \to \infty$ yields the following: For $t \geq 0$,

$$\limsup_{t\to\infty} \frac{1}{t}\mathbb{E}\big[\xi(t)\big] = \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}\big[\Xi(t)\big] - \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}\big[y'X(t)\big] + \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}\big[y'z(t, W(t))\big] \tag{C.31}$$

$$= \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}\big[\Xi(t)\big] + y'\zeta, \tag{C.32}$$

where (C.31) follows from (5.12) and (C.32) follows from the fact that X is a Brownian motion with drift vector $-\zeta$, and z is bounded. □

Proof of Lemma 5.5. The desired property can be deduced from the optimality conditions of the SPP (4.2). To that end, we let $\hat{y} \in \mathbb{R}^I$, $\hat{\rho} \in \mathbb{R}^K_+$, $\hat{\xi} \in \mathbb{R}^I_+$, and $\hat{\phi} \in \mathbb{R}^J_+$ denote the Lagrange multipliers associated with the flow-balance constraint, capacity constraint, nonnegativity constraint for o, and nonnegativity constraint for n, respectively. Then, the Lagrangian is given as follows:

$$\mathcal{L}(\hat{\xi}, \hat{\phi}, \hat{\rho}, \hat{y}) = \min_{\mu\in\mathcal{M}, n, o} \sum_{j\in\mathcal{J}} n_j\mu_j\upsilon_j(\mu_j, 0) + \sum_{i\in\mathcal{I}} \nu_i o_i + \hat{y}'(R(\mu, 0)n + o - \lambda) + \hat{\rho}'(An - e) - \hat{\xi}'o - \hat{\phi}'n.$$

Differentiating the Lagrangian with respect to the service rate μ_j and evaluating it at the optimal

solution $(\mu^\star, n^\star, o^\star)$ yields the following: For basic activities $j = 1, \ldots, J_b$,

$$n_j^\star\big(\upsilon_j(\mu_j^\star, 0) + \mu_j^\star \frac{\partial \upsilon_j}{\partial \mu_j}(\mu_j^\star, 0)\big) \; + \; \sum_{i=1}^{I} \hat{y}_i\big((\mathbf{1}_{\{i(j)=i\}} - P_{ji}(\mu_j^\star, 0))n_j^\star - n_j^\star \mu_j^\star \frac{\partial P_{ji}}{\partial \mu_j}(\mu_j^\star, 0)\big) \; = \; 0. \tag{C.33}$$

Substituting (4.14) and (4.15) into (C.33) yields the following: For $j = 1, \ldots, J_b$,

$$b_j \; + \; (C'\hat{y})_j \; = \; 0. \tag{C.34}$$

Moreover, it follows from the first-order optimality condition with respect to n_j for the basic activities, combined with complementary slackness, that $(\hat{y}, \hat{\rho})$ satisfies (5.1). By Proposition 5.3, the pair (y, ρ) that solves (5.1) is determined up to a one-dimensional space. Then, because $m'R(\mu^\star, 0) \; = \; g'H$, we deduce that the pair $(\hat{y} + \alpha m, \hat{\rho} - \alpha g)$, for scalar α, as long as $\hat{\rho} - \alpha g \in \mathbb{R}_+^{J-J_b+K}$ also solves (5.1). Substituting $y = \hat{y} + \alpha m$ into (5.6) yields

$$\beta \; = \; b \; + \; C'(\hat{y} + \alpha m) \; = \; b \; + \; C'\hat{y} \; + \; \alpha C'm. \tag{C.35}$$

Then, by substituting (C.34) into (C.35) we have that for basic activities $j = 1, \ldots, J_b, \beta_j = \alpha(C'm)_j$. Finally, it follows from the definition of b and C that all terms corresponding to nonbasic activities are equal to zero. Therefore, for nonbasic activities $j = J_b + 1, \ldots, J$, we have that $\beta_j = 0$. □

Proof of Proposition 6.2. Consider a barrier policy (L, U, z, θ) with a lower barrier at l and an upper barrier at u, and assume that $\gamma \in \mathbb{R}$ and $f \in C^2([l, u])$ jointly satisfy (6.3)–(6.4). A routine application of Itô's lemma on the workload process in (5.13) yields

$$\begin{aligned}
\mathbb{E}[f(W(t))] - f(W(0)) \; &= \; \mathbb{E}\Big[\int_0^t \Gamma_{z,\theta} f(W(s))ds\Big] + \mathbb{E}\Big[\int_0^t f'(W(s))dL(s)\Big] - \mathbb{E}\Big[\int_0^t f'(W(s))dU(s)\Big] \\
&= \; \mathbb{E}\Big[\int_0^t \Gamma_{z,\theta} f(W(s))ds\Big] - \mathbb{E}\Big[\int_0^t \kappa_L dL(s)\Big] - \mathbb{E}\Big[\int_0^t \kappa_U dU(s)\Big] \quad \text{(C.36)} \\
&= \; \gamma t - \mathbb{E}\Big[\int_0^t c(\theta(W(s)))ds + \int_0^t h'z(W(s))ds\Big] - \mathbb{E}\Big[\kappa_L L(t)\Big] - \mathbb{E}\Big[\kappa_U U(t)\Big] \quad \text{(C.37)} \\
&= \; \gamma t \; - \; \mathbb{E}\Big[\int_0^t c(\theta(W(s)))ds + \int_0^t h'z(W(s))ds \; + \; \kappa_L L(t) \; + \; \kappa_U U(t)\Big], \quad \text{(C.38)}
\end{aligned}$$

where (C.36) follows from (6.1) and (6.4), (C.37) follows from (6.3), and (C.38) follows by

rearranging the term. Dividing both sides of (C.38) by t yields

$$\frac{1}{t}\mathbb{E}[f(W(t))] - \frac{1}{t}f(W(0)) = \gamma - \frac{1}{t}\mathbb{E}\Big[\int_0^t c(\theta(W(s)))ds + \int_0^t h'z(W(s))ds + \kappa_L L(t) + \kappa_U U(t)\Big]. \tag{C.39}$$

As $t \to \infty$, both terms on the left-hand side of (C.39) vanish because f is a continuous function on a compact interval and $W(t) \in [l, u]$ for $t \geq 0$. Substituting this into (C.39), taking the lim as $t \to \infty$, and rearranging the terms yields the desired result. □

Proof of Lemma 6.3. We prove each statement separately.

Part (i). By (5.7), $c(x)$ is equal to the optimal objective function value of a linear program parameterized by $x \in \mathbb{R}$. It follows from standard results in parametric linear programming (see, e.g., Bertsimas and Tsitsiklis (1997, Section 5.2)) that c is convex, continuous, piecewise linear. Then, since ψ is the negative of the convex conjugate of c, it follows from Boyd and Vandenberghe (2004, Section 3.3.1) that ψ is concave and continuous. Moreover, by Rockafellar (2015, Theorem 19.2), ψ is piecewise linear.

Part (ii). Substituting (5.9) into (6.9) yields

$$\phi(w, y) = \min\big\{h'z - d'zy : z \in \mathbb{R}^I_+ : m'z = w,\ \Lambda z \leq \delta\big\}, \quad w \in [\underline{w}, \bar{w}], y \in \mathbb{R}. \tag{C.40}$$

For each fixed $y \in \mathbb{R}$, the mapping $w \mapsto \phi(w, y)$ is equal to the optimal objective function value of a linear program with a compact convex feasible set parameterized by w. It follows from standard results in parametric linear programming (see, e.g., Bertsimas and Tsitsiklis (1997, Section 5.2)) that $w \mapsto \phi(w, y)$ is convex, continuous, and piecewise linear. To establish the Lipschitz property with respect to y, note that for any feasible z, the objective function in (C.40) is affine in y with slope $-d'z$. Because the feasible set is uniformly bounded for $w \in [\underline{w}, \bar{w}]$, there exists $L > 0$ such that $|d'z| \leq L$ for $w \in [\underline{w}, \bar{w}]$ and $z \in \mathcal{A}(w)$. Consequently, for any $y_1, y_2 \in \mathbb{R}$ and $w \in [\underline{w}, \bar{w}]$, we have that $|\phi(w, y_1) - \phi(w, y_2)| \leq L|y_1 - y_2|$, which yields the desired result. □

Proof of Lemma 6.4. Suppose (l, u, γ, v) is a solution to the Bellman equation (6.10)–(6.12). First, we show that $\gamma \in [\underline{\gamma}, \infty)$. Suppose, to the contrary, that $\gamma \in (-\infty, \underline{\gamma})$. Then, it follows from (6.13) that

$$\gamma + \mu\kappa_L - \psi(-\kappa_L) - \phi(w, -\kappa_L) < 0, \quad w \in [\underline{w}, \bar{w}]. \tag{C.41}$$

It follows from (6.10) and (C.41) that for any $w \in [\underline{w}, \bar{w}]$ at which $v(w) = -\kappa_L$, we have that $v'(w) < 0$. Then, since $v(l) = -\kappa_L$, and consequently, $v'(l) < 0$, and v is continuously differentiable,

it follows from (6.10) that $v(l+\epsilon) < -\kappa_L$ for some $\epsilon > 0$. However, since $v(u) = \kappa_U > -\kappa_L$ and v is continuous, there must exists $w \in (l+\epsilon, \bar{w})$ such that $v(w) = -\kappa_L$ and $v'(w) \geq 0$. However, this contradicts the previous conclusion that $v'(w) < 0$. This completes the proof of the first claim.

Next, we assume $l > \underline{w}$ and show that the mapping $w \mapsto \phi(w, -\kappa_L)$ is nonincreasing on $[\underline{w}, l]$. Since $w \mapsto \phi(w, -\kappa_L)$ is convex, it suffices to show that it is nonincreasing in a neighborhood of l. We argue this by contradiction. Suppose to the contrary that $w \mapsto \phi(w, -\kappa_L)$ is not nonincreasing in a neighborhood of l. It follows the fact that $w \mapsto \phi(w, -\kappa_L)$ is convex, continuous, and piecewise linear (by Lemma 6.3) that it is strictly increasing in a neighborhood to the right of l. Since $w \mapsto \phi(w, -\kappa_L)$ is continuous, piecewise linear, and convex (by Lemma 6.3), it follows that it is strictly increasing on $[l, u]$. It then follows from (6.10)–(6.12) that v is decreasing in a neighborhood of l, and for any $w \in (l, u]$ for which $v(w) = -\kappa_L$, we must have $v'(w) < 0$. However, since $v(u) = \kappa_U > -\kappa_L$ and v is continuously differentiable, there must exist $w \in (l, u)$ such that $v(w) = -\kappa_L$ and $v'(w) \geq 0$, which contradicts the earlier conclusion. Therefore, the mapping $w \mapsto \phi(w, -\kappa_L)$ is nonincreasing in a neighborhood to the right of l, and the desired results follows from the convexity of $w \mapsto \phi(w, -\kappa_L)$. Finally, substituting $v(l) = -\kappa_L$ into (6.10) yields (6.14), which completes the proof. □

Proof of Lemma 6.5. We begin by proving that the mapping $\gamma \mapsto {}_\gamma$ is nonincreasing on $[\underline{\gamma}, \infty)$. In other words, given $\gamma_1, \gamma_2 \in [\underline{\gamma}, \infty)$ with $\gamma_1 \geq \gamma_2$, we have that $l_{\gamma_1} \leq l_{\gamma_2}$. For ease of exposition, we define the function $F : [\underline{w}, \bar{w}] \to \mathbb{R}$ as follows:

$$F(w) \coloneqq -\mu\kappa_L + \psi(-\kappa_L) + \phi(w, -\kappa_L), \quad w \in [\underline{w}, \bar{w}]. \tag{C.42}$$

By construction, $l_\gamma \geq \underline{w}$ for $\gamma \in [\underline{\gamma}, \infty)$, and $l_\gamma = \underline{w}$ whenever $\mathcal{L}_\gamma = \varnothing$. Hence if $\mathcal{L}_{\gamma_2} = \varnothing$, then $l_{\gamma_2} = \underline{w}$ and the result is trivial. Now suppose $\mathcal{L}_{\gamma_2} \neq \varnothing$. For $\gamma \in [\underline{\gamma}, \infty)$, if $l_\gamma > \underline{w}$, then by definition, l_γ is the minimal $w \in [\underline{w}, \bar{w}]$ such that $F(w) = \gamma$ and $w \mapsto \phi(w, -\kappa_L)$ is nonincreasing on $[\underline{w}, l_\gamma)$. By Lemma 6.3, $w \mapsto \phi(w, -\kappa_L)$ is convex and piecewise linear, and by (C.42), F is nonincreasing in any region where $w \mapsto \phi(w, -\kappa_L)$ is nonincreasing. Because F is nonincreasing wherever $w \mapsto \phi(w, -\kappa_L)$ is nonincreasing, the existence of w at which $F(w) = \gamma_2$ and $w \mapsto \phi(w, -\kappa_L)$ is nonincreasing implies the existence of such w for any $\gamma_1 \leq \gamma_2$. Hence $\mathcal{L}_{\gamma_1} \neq \varnothing$. Moreover, because F is nonincreasing wherever $w \mapsto \phi(w, -\kappa_L)$ is nonincreasing, increasing γ from γ_1 to γ_2 can only reduce (or leave unchanged) the minimal w solving $F(w) = \gamma$ with $w \mapsto \phi(w, -\kappa_L)$ being nonincreasing. Hence $l_{\gamma_2} \leq l_{\gamma_1}$.

We now turn to the continuity of the mapping $\gamma \mapsto l_\gamma$ on $[\underline{\gamma}, \infty)$. If $l_{\underline{\gamma}} = \underline{w}$, the result is trivial. Suppose $l_{\underline{\gamma}} > \underline{w}$. It follows from the fact that F is convex (by (C.42) and Lemma 6.3) and $l_{\underline{\gamma}}$ is the smallest w such that $F(w) = \underline{\gamma}$ and $w \mapsto \phi(w, -\kappa_L)$ is nonincreasing on $[\underline{w}, w]$ that F is strictly decreasing on $[\underline{w}, l_{\underline{\gamma}}]$. Therefore, F is a continuous bijection from $[\underline{w}, l_{\underline{\gamma}}]$ onto $[\underline{\gamma}, F(\underline{w})]$. Let

$\gamma_0 \coloneqq F(\underline{w})$. Then, for any $\gamma \in [\underline{\gamma}, \gamma_0]$, the set $\mathcal{L}_\gamma$ is a singleton, and l_γ is given by the inverse of F restricted to $[\underline{w}, l_{\underline{\gamma}}]$. Hence the mapping $\gamma \mapsto l_\gamma$ is continuous on $[\underline{\gamma}, \gamma_0]$ with $l_{\gamma_0} = \underline{w}$. Next, note that the mapping $\gamma \to l_\gamma$ is nonincreasing on (γ_0, ∞). Therefore, $l_\gamma = \underline{w}$ for $\gamma \in (\gamma_0, \infty)$, which completes the proof. □

Proof of Lemma 6.6.

The continuity of V follows from Berge's Maximum Theorem (see, e.g., Sundaram (1996, Theorem 9.14)) and that the mapping $(\gamma, w) \mapsto v_\gamma(w)$ is jointly continuous (by Lemma C.2), the mapping $\gamma \mapsto l_\gamma$ is continuous (by Lemma 6.5), and for each $\gamma \in [\underline{\gamma}, \infty)$, the set $[l_\gamma, \bar{w}]$ is compact.

Next, we show that V is strictly increasing. To that end, consider $\gamma_1, \gamma_2 \in [\underline{\gamma}, \infty)$ with $\gamma_1 < \gamma_2$. Let $\hat{w} \in [l_{\gamma_1}, \bar{w}]$ denote a point attaining the maximum of v_{γ_1}, so that $V(\gamma_1) = v_{\gamma_1}(\hat{w})$. Since $\gamma \mapsto l_\gamma$ is nonincreasing (by Lemma 6.5), it follows that $[l_{\gamma_1}, \bar{w}] \subseteq [l_{\gamma_2}, \bar{w}]$, and thus $\hat{w} \in [l_{\gamma_2}, \bar{w}]$. Lemma C.2 implies that $\gamma \mapsto v_\gamma(\hat{w})$ is strictly increasing, and hence, $v_{\gamma_2}(\hat{w}) > v_{\gamma_1}(\hat{w})$. Therefore,

$$V(\gamma_2) \geq v_{\gamma_2}(\hat{w}) > v_{\gamma_1}(\hat{w}) = V(\gamma_1),$$

which establishes that V is strictly increasing as desired.

It remains to show that $V(\underline{\gamma}) < \kappa_U$ and $V(\gamma) \to \infty$ as $\gamma \to \infty$. The first claim immediately follows from (6.18) and Lemma C.2. We argue the second claim by contradiction. Suppose to the contrary that V is bounded. Then, it follows from the strict monotonicity of v_γ and the fact that V is bounded that v_γ is bounded (above and below). Then, it follows from (6.16) that $v'_\gamma(w) \to \infty$ as $\gamma \to \infty$ for $w \in [\underline{w}, \bar{w}]$, which contradicts the bounded of v_γ and completes the proof. □

Proof of Theorem 6.7.

It follows from (6.16)–(6.17), (6.19), and continuity of $v^\star$ that $v^\star$ satisfies

$$\frac{1}{2}\sigma^2 v'(w) = \gamma^\star - \mu v(w) - \psi(v(w)) - \phi(w, v(w)), \quad w \in [l^\star, u^\star]$$

subject to the boundary condition $v^\star(l^\star) = -\kappa_L$, and the terminal condition $v^\star(u^\star) = \kappa_U$. Moreover, $v^\star(w) < \kappa_U$ for $w \in [l^\star, u^\star)$. It remains to verify (6.12). If $l^\star > \underline{w}$, by construction (see (6.15)) we have that $v'(l^\star) = 0$. Similarly, if $u^\star < \bar{w}$, it follows from the minimality of $u^\star$ that $v'(u^\star) = 0$. To be specific, if $v'(u^\star) \neq 0$, it follows from the continuous differentiability of $v_{\gamma^\star}$ that there exists $w \neq u^\star$ such that $v_{\gamma^\star}(w) > \kappa_U$, which violates the minimality of $\gamma^\star$. Therefore, $(l^\star, u^\star, \gamma^\star, v^\star)$ satisfies the Bellman equation (6.10)–(6.12).

Next, we show that if $u^\star < \bar{w}$, then the mapping $w \mapsto \phi(w, \kappa_U)$ is nondecreasing on $[u^\star, \bar{w}]$. It follows from (6.10), the continuity of $v^\star$ and $(v^\star)'$, continuity and almost everywhere differentiability of ψ and ϕ, and $(v^\star)'(u^\star) = 0$ that $(v^\star)''$ is well-defined almost everywhere in a neighborhood of

$u^\star$. It follows from the minimality of $u^\star$ and $(v^\star)'(u^\star) = 0$ that $(v^\star)'(w) > 0$ and $(v^\star)''(w) < 0$ for $w \in (u^\star - \delta, u^\star)$ for some $\delta > 0$. If $\delta > 0$ is chosen to be sufficiently small, ϕ and ψ are affine on $(u^\star - \delta, u^\star)$. Then, by taking derivative from both sides of (6.10) and using $(v^\star)'(u^\star) = 0$ and $(v^\star)''(w) < 0$ for $w \in (u^\star - \delta, u^\star)$, we obtain $\partial\phi(w, \kappa_U)/\partial w \geq 0$ for $w \in (u^\star - \delta, u^\star)$. In other words, $w \mapsto \phi(w, \kappa_U)$ is nondecreasing on $(u^\star - \delta, u^\star)$. The desired result then follows from the convexity of $w \mapsto \phi(w, \kappa_U)$ (by Lemma 6.3). □

Proof of Theorem 6.10.

Let $(l^\star, u^\star, \gamma^\star, v^\star)$ be the solution to the Bellman equation (6.10)–(6.12) constructed in (6.19). Consider the barrier policy $(L^\star, U^\star, z^\star, \theta^\star)$ with a lower barrier at $l^\star$, an upper barrier at $u^\star$, effective drift rate function $\theta^\star$ defined by (6.21), and workload configuration function $z^\star$ defined by (6.22). It follows from Proposition 6.2 and Corollary 6.8 that the barrier policy $(L^\star, U^\star, z^\star, \theta^\star)$ achieves a long-run average expected cost of $\gamma^\star$. To complete the proof, it remains to show that $\gamma^\star$ is a lower bound on the long-run average expected cost of any admissible policy (L, U, z, θ) to the EWF (5.17). The remainder of the proof consists of two parts. In the first part, we extend the solution of the Bellman equation to $[\underline{w}, \bar{w}]$ and establish an inequality involving $\gamma^\star$. In the second part, we use the results from the first part, along with Itô's Lemma, to prove that no admissible policy can achieve a long-run average expected cost lower than $\gamma^\star$.

We begin by extending the relative value function $f^\star : [l^\star, u^\star] \to \mathbb{R}$ defined in (6.20) to the entire interval $[\underline{w}, \bar{w}]$ as follows:

$$f^\star(w) := \begin{cases} \kappa_L \left(l^\star - w\right), & w \in [\underline{w}, l^\star), \\ \displaystyle\int_{l^\star}^{w} v^\star(x)dx, & w \in [l^\star, u^\star], \\ \displaystyle\int_{l^\star}^{u^\star} v^\star(x)dx \, + \, \kappa_U(w \, - \, u^\star), & w \in (u^\star, \bar{w}]. \end{cases} \tag{C.43}$$

We now verify some properties of this extended function. First, since $(l^\star, u^\star, \gamma^\star, v^\star)$ solves the Bellman equation (6.10)–(6.12), it follows from (C.43) that $f^\star \in C^2([\underline{w}, \bar{w}])$. Second, we have $(f^\star)' \equiv -\kappa_L$ on $[\underline{w}, l^\star]$, $(f^\star)' \equiv v^\star$ on $[l^\star, u^\star]$, and $(f^\star)' \equiv \kappa_U$ on $[u^\star, \bar{w}]$. Therefore, by Theorem 6.7, $-\kappa_L \leq (f^\star)'(w) \leq \kappa_U$ for $w \in [\underline{w}, \bar{w}]$. We prove the main result in two steps. In Step 1, we show that

$$\frac{1}{2}\sigma^2(f^\star)''(w) \, + \, \mu(f^\star)'(w) \, + \, \psi((f^\star)'(w)) \, + \, \phi(w, (f^\star)'(w)) \, \geq \, \gamma^\star, \quad w \in [\underline{w}, \bar{w}]. \tag{C.44}$$

In Step 2, we show that the long-run average cost of any arbitrary admissible policy (L, U, z, θ) is greater than or equal to the long-run average cost of our candidate policy.

Step 1. We verify (C.44) on the three sets $[l^\star, u^\star]$, $[\underline{w}, l^\star)$, and $(u^\star, \bar{w}]$, separately. First, we consider the set $[l^\star, u^\star]$. For $w \in [l^\star, u^\star]$, (C.44) holds with equality since $(l^\star, u^\star, \gamma^\star, f^\star)$ solves the Bellman equation (6.5)–(6.7); see Corollary 6.11. Next, we consider the set $[\underline{w}, l^\star]$. The result is trivial if $l^\star = \underline{w}$. If $l^\star > \underline{w}$, then it follows from the fact that (C.44) holds with equality for $w = l^\star$ and the mapping $w \mapsto \phi(w, -\kappa_L)$ is nonincreasing at $w = l^\star$ (by (6.15)) and convex (by Lemma 6.3) that (C.44) holds for $w \in [\underline{w}, l^\star]$. Finally, we consider the set $(u^\star, \bar{w}]$. The result is trivial if $u^\star = \bar{w}$. If $u^\star < \bar{w}$, then it follows from the fact that (C.44) holds with equality for $w = u^\star$ and the mapping $w \mapsto \phi(w, \kappa_U)$ is nondecreasing at $w = u^\star$ (by Theorem 6.7) and is convex (by Lemma 6.3) that (C.44) holds for $w \in [u^\star, \bar{w}]$.

Step 2. Let (L, U, z, θ) be an arbitrary admissible policy to the EWF (5.17) and let W be the resulting workload process defined in (5.13). A routine application of Itô's lemma yields the following: For $t \geq 0$,

$$\mathbb{E}\Big[f^\star(W(t)) - f^\star(W(0))\Big] = \mathbb{E}\Big[\int_0^t \Gamma_{z,\theta} f^\star(W(s))ds\Big] + \mathbb{E}\Big[\int_0^t (f^\star)'(W(s))\, d(L(s) - U(s))\Big], \tag{C.45}$$

where the differential operator $\Gamma_{z,\theta}$ is given by (6.2). Next, we derive a lower bound for each of the terms on the right-hand side of (C.45). For the first term, we use (6.8)–(6.9) to obtain a lower bound for $\Gamma_{z,\theta} f^\star(W(s))$ for $s \geq 0$. Then, we integrate both sides of the inequality over $[0, t]$, take expectation, and rearrange the terms, which yields

$$\mathbb{E}\Big[\int_0^t \Gamma_{z,\theta} f^\star(W(s))ds\Big] \geq \gamma^\star t - \mathbb{E}\Big[\int_0^t h'z(W(s))ds\Big] - \mathbb{E}\Big[\int_0^t c(\theta(s))ds\Big] \tag{C.46}$$

for $t \geq 0$. For the second term, note that L and U are nondecreasing and $-\kappa_L \leq (f^\star)'(w) \leq \kappa_U$ for $w \in [\underline{w}, \bar{w}]$. Therefore, for $t \geq 0$,

$$\mathbb{E}\Big[\int_0^t (f^\star)'(W(s))\, dL(s)\Big] \geq -\mathbb{E}\Big[\kappa_L L(t)\Big] \quad \text{and} \quad \mathbb{E}\Big[\int_0^t (f^\star)'(W(s))\, dU(s)\Big] \leq \mathbb{E}\Big[\kappa_U U(t)\Big]. \tag{C.47}$$

Substituting (C.46)–(C.47) into (C.45) gives

$$\mathbb{E}\Big[f^\star(W(t)) - f^\star(W(0))\Big] \geq \gamma^\star t - \mathbb{E}\Big[\int_0^t c(\theta(s))ds + \int_0^t h'z(W(s))ds + \kappa_L L(t) + \kappa_U U(t)\Big] \tag{C.48}$$

for $t \geq 0$. We divide both sides of (C.48) by t and take lim inf as $t \to \infty$. The terms on the left-hand

side vanish because $(f^\star)'$ is bounded and $W(t) \in [\underline{w}, \bar{w}]$ for $t \geq 0$, which yields

$$\liminf_{t\to\infty} \frac{1}{t}\mathbb{E}\Big[\int_0^t c(\theta(s))ds + \int_0^t h'z(W(s))ds + \kappa_L L(t) + \kappa_U U(t)\Big] \geq \gamma^\star. \qquad \square$$

C.3 Auxiliary Results

This section presents a number of auxiliary lemmas that are used in the proofs of the main results. Throughout this section, we retain the notation and assumptions laid out in Chapters 5 and 6. Given $\gamma, \kappa \in \mathbb{R}$, we consider the following differential equation:

$$\frac{1}{2}\sigma^2 v'(w) = \gamma - \mu v(w) - \psi(v(w)) - \phi(w, v(w)), \quad w \in [l, \bar{w}], \tag{C.49}$$

subject to the boundary condition

$$v(l) = \kappa. \tag{C.50}$$

Lemma C.3. *For given parameters $\gamma, \kappa \in \mathbb{R}$, let $v_{\gamma,\kappa} \in C^1([l, \bar{w}])$ denote the unique solution to (C.49)–(C.50). Suppose $\gamma_1 > \gamma_2$ and $\kappa_1 > \kappa_2$. Then, $v_{\gamma_1,\kappa_1}(w) > v_{\gamma_2,\kappa_2}(w)$ for $w \in [l, \bar{w}]$.*

Proof. Proof. We argue by contradiction. Suppose the conclusion fails. Then, there exists $w \in (l, \bar{w}]$ such that $v_{\gamma_1,\kappa_1}(w) \leq v_{\gamma_2,\kappa_2}(w)$. Let

$$w^\star := \min\big\{w : v_{\gamma_1,\kappa_1}(w) \leq v_{\gamma_2,\kappa_2}(w), w \in (l, \bar{w}]\big\}.$$

It follows from $v_{\gamma_1,\kappa_1}(l) = \kappa_1 > \kappa_2 = v_{\gamma_2,\kappa_2}(l)$ and the fact that v_{γ_1,κ_1} and v_{γ_2,κ_2} are continuous that

$$v_{\gamma_1,\kappa_1}(w^\star) = v_{\gamma_2,\kappa_2}(w^\star) \quad \text{and} \quad v_{\gamma_1,\kappa_1}(w) > v_{\gamma_2,\kappa_2}(w), \quad w \in [l, w^\star).$$

It follows that $v'_{\gamma_1,\kappa_1}(w^\star) \leq v'_{\gamma_2,\kappa_2}(w^\star)$. However, evaluating both sides of (C.49) at $w^\star$ and using the fact that $\gamma_1 > \gamma_2$ yields $v'_{\gamma_1,\kappa_1}(w^\star) > v'_{\gamma_2,\kappa_2}(w^\star)$, which contradicts the earlier conclusion. Therefore, $v_{\gamma_1,\kappa_1}(w) > v_{\gamma_2,\kappa_2}(w)$ for $w \in [l, \bar{w}]$. $\square$ $\square$

Lemma C.4. *Fix $\gamma, \kappa \in \mathbb{R}$ and let $v \in C^1([l, \bar{w}])$ denote the unique solution to (C.49)–(C.50). Suppose there exists $\delta > 0$ such that*

$$\gamma - \mu\kappa - \psi(\kappa) - \phi(w, \kappa) < 0, \quad w \in (l, l + \delta), \tag{C.51}$$

and that $w \mapsto \phi(w, \kappa)$ *is nondecreasing on* $(l, \bar{w}]$. *Then,* $v(w) \leq \kappa$ *for* $w \in [l, \bar{w}]$.

Proof. We first establish that $v(w) < \kappa$ for $w \in (l, l + \epsilon)$ for some $\epsilon > 0$. It follows from (C.51) and continuity of $w \mapsto \phi(w, \kappa)$ (by Lemma 6.3) that

$$\gamma - \mu\kappa - \psi(\kappa) - \phi(l, \kappa) \leq 0. \tag{C.52}$$

First, suppose (C.52) holds as strict inequality. It follows from (C.49) that $v'(l) < 0$, which by the continuous differentiability of v implies $v'(w) < 0$ for $w \in [l, l + \epsilon)$ for some $\epsilon > 0$. Consequently, $v(w) < \kappa$ for $w \in [l, l + \epsilon)$ as desired. Next, suppose (C.52) holds as equality. It follows from (C.49) that $v'(l) = 0$. Moreover, it follows from (C.51)–(C.52) and the fact that $w \mapsto \phi(w, \kappa)$ is convex and piecewise linear (by Lemma 6.3) that $w \mapsto \phi(w, \kappa)$ is strictly increasing and affine on $(l, l + \hat{\epsilon})$ for some $\hat{\epsilon} > 0$. It then follows from (C.49), continuity and almost everywhere differentiability of ψ and ϕ, continuity of v and v', and $v'(l) = 0$ that v'' is well-defined almost everywhere and negative on $[l, l + \epsilon)$ for some $\epsilon > 0$. Therefore, $v'(w) < 0$ for $w \in (l, l + \epsilon)$, and consequently, $v(w) < \kappa$ for $w \in (l, l + \epsilon)$ as desired.

To complete the proof, we argue by contradiction that $v(w) \leq \kappa$ for $w \in [l, \bar{w}]$. Suppose to the contrary that there exists $w_0 \in (l, \bar{w}]$ such that $v(w_0) > \kappa$. Then, since $v(w) < \kappa$ on $(l, l + \epsilon)$ for some $\epsilon > 0$ as established earlier, there exists $w^\star \in (l + \epsilon, w_0)$ such that $v(w^\star) = \kappa$ and $v'(w^\star) \geq 0$. However, by (C.49) and the fact that $w \mapsto \phi(w, \kappa)$ is nondecreasing on $(l, \bar{w}]$, we have that

$$\frac{1}{2}\sigma^2 v'(w^\star) = \gamma - \mu\kappa - \psi(\kappa) - \phi(w^\star, \kappa) \leq \gamma - \mu\kappa - \psi(\kappa) - \phi(l + (\delta \wedge \epsilon)/2, \kappa) < 0,$$

where the last inequality follows from (C.51). This contradicts the earlier conclusion that $v'(w^\star) \geq 0$ and completes the proof. □ □

Lemma C.5. *Fix* $\gamma, \kappa \in \mathbb{R}$ *and let* $v \in C^1([l, \bar{w}])$ *denote the unique solution to (C.49)–(C.50). Suppose there exists* $\delta > 0$ *such that*

$$\gamma - \mu\kappa - \psi(\kappa) - \phi(w, \kappa) > 0, \quad w \in (l, l + \delta), \tag{C.53}$$

and that $w \mapsto \phi(w, \kappa)$ *is strictly decreasing on* $(l, l + \delta)$. *Then,* $v(w) > \kappa$ *for* $w \in (l, l + \delta]$.

Proof. It follows from (C.53) and continuity of $w \mapsto \phi(w, \kappa)$ (by Lemma 6.3) that

$$\gamma - \mu\kappa - \psi(\kappa) - \phi(l, \kappa) \geq 0. \tag{C.54}$$

First, suppose (C.54) holds as strict inequality. It follows from (C.53)–(C.54), continuity and piecewise linearity of ϕ and ψ (by Lemma 6.3), continuity of v and v', and the fact that $w \mapsto \phi(w, \kappa)$

is convex (by Lemma 6.3) and nonincreasing on $[l, l+\delta)$ that there exists $\epsilon > 0$ such that if $v(w) \in [\kappa, \kappa+\epsilon)$ and $w \in [l, l+\delta)$, then $v'(w) > 0$. This then implies that $v(w) > v(l) = \kappa$ for $w \in (l, l+\delta]$ as desired.

Next, suppose (C.54) holds as equality. It follows from (C.49) that $v'(l) = 0$. The mapping $w \mapsto \phi(w, \kappa)$ is strictly decreasing and piecewise (by Lemma 6.3) on $(l, l+\delta)$. Therefore, it follows from (C.49), continuity and almost everywhere differentiability of ψ and ϕ, continuity of v and v', and $v'(l) = 0$ that v'' is well-defined almost everywhere and positive on $[l, l+\epsilon)$ for some $\epsilon > 0$. Therefore, $v'(w) > 0$ for $w \in (l, l+\epsilon)$, and consequently, $v(w) > \kappa$ for $w \in (l, l+\epsilon)$. Then, as discussed earlier, one can argue that there exists $\hat{\epsilon} > 0$ such that if $v(w) \in (\kappa, \kappa+\hat{\epsilon})$ and $w \in (l, l+\delta)$, then $v'(w) > 0$. This then implies that $v(w) > v(l) = \kappa$ for $w \in (l, l+\delta]$ as desired. □ □

Appendix D

Supplementary Material for the Simulation Study

This section presents supplementary graphs and tables for the simulation study discussed in Chapter 8. These materials provide additional detail on model primitives, cost functions, transition probabilities, and sensitivity analyses that underlie the numerical results discussed in Chapter 8.

Table D.1 reports the parameters of the transition kernels and the probability that a customer does not reach the desired service outcome after an episode of service. Tables D.2 and D.3 report the parameters of the workload formulation for the base case discussed in Section 8.1. Tables D.2 reports the parameters associated with the buffers and Table D.3 reports the parameters associated with the pushing controls.

Table D.1: Parameters of the transition kernels and the probability that a customer does not reach the desired service outcome after an episode of service. Entries marked "–" denote parameters that are not applicable.

j	1	2	3	4	5	6	7	8	9	10
a_{1j}	2.01	2.01	0.04	0.04	2.01	2.01	2.01	2.01	2.01	2.01
b_{1j}	-66.54	-59.89	-15.96	-17.73	-59.89	-66.54	-59.89	-66.54	-66.54	-59.89
c_{1j}	0.30	0.30	0.30	0.30	0.24	0.24	0.24	0.24	0.24	0.24
a_{2j}	–	–	4.62	4.62	–	–	–	–	–	–
b_{2j}	–	–	-252.40	-252.40	–	–	–	–	–	–
c_{2j}	–	–	0.80	0.80	–	–	–	–	–	–

Figure D.1 depicts the service cost function of activities $j = 1, \ldots, 4$ (i.e., those serving classes 1 and 2) evaluated at a waiting time of zero. The service cost is increasing and strictly convex in the expected service time $1/\mu_j$, implying that progressively larger costs are incurred as treatment intensity decreases. The function is normalized so that its maximum value at zero waiting time

Table D.2: Parameters associated with buffers.

i	1	2	3	4	5
m_i	0.30	0.24	0.26	0.32	0.27
ν_i	6.00	16.00	4.00	8.00	4.00
π_i	3.00	8.00	2.00	4.00	2.00
y_i	-2.04	-1.67	-2.20	-2.58	-2.40
h_i	0.26	0.26	0.27	0.33	0.28
d_i	-0.018	-0.025	0.00	0.00	0.00

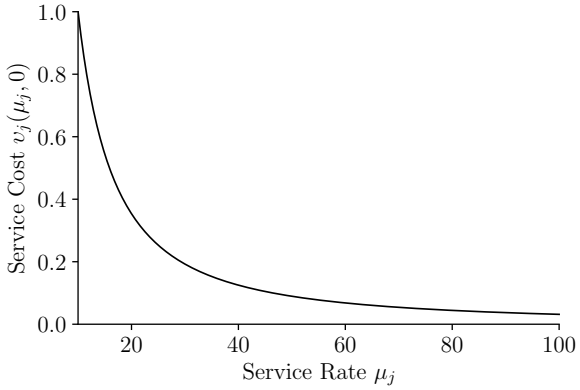


Figure D.1: Service cost function of activities $j = 1, \ldots, 4$ evaluated at a waiting time of zero for the numerical example presented in Section 8.

Table D.3: Parameters associated with pushing controls.

k	1	2	3	4	5	6
g_k	0.53	0.47	0.09	-0.20	-0.27	0.001
ρ_k	2.42	2.08	0.89	1.29	4.13	0.00

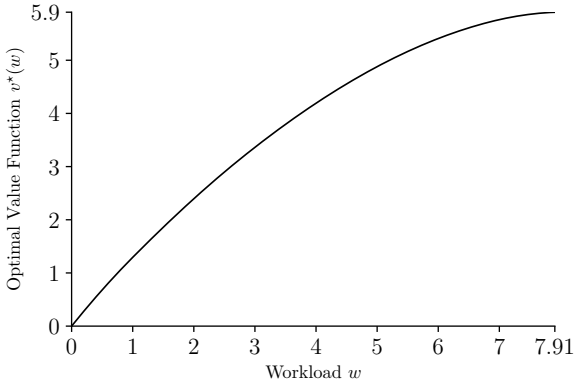


Figure D.2: The optimal value function $v^\star$ of the workload formulation for the base case of the numerical example presented in Section 8.

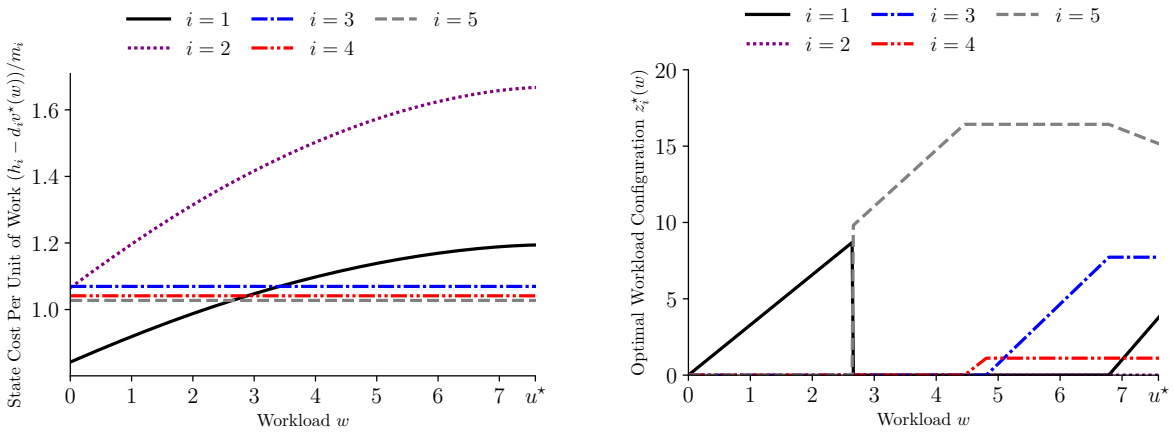
equals one, which facilitates comparison across the different cost components in the model.

Figure D.2 depicts the optimal value function of the workload formulation for the base case. The value function is increasing and concave in workload, reflecting the increasing marginal burden of congestion together with the ability of the dynamic policy to partially mitigate congestion through service rate adjustments and workload regulation. The slope of the value function corresponds to the marginal cost of additional workload and plays a central role in determining the optimal barrier levels and workload configuration.

Figures D.3–D.9 plot the state cost per unit of work and the corresponding optimal workload configuration for each class under a range of alternative parameter settings. In each figure, exactly one model parameter—such as the sensitivity of service outcomes and transitions to waiting times, the waiting costs, the abandonment rates, or the abandonment costs—is varied relative to the base case. All remaining parameters are held fixed at their base-case values, as specified in Section 8.1. These figures allow the reader to visualize how changes in individual parameters affect the structure of the workload formulation solution.

Tables D.5, D.6, D.7, and D.8 compare the long-run average cost of the various policies as we vary the sensitivity of the service outcomes and transitions to waiting times, abandonment rates, and

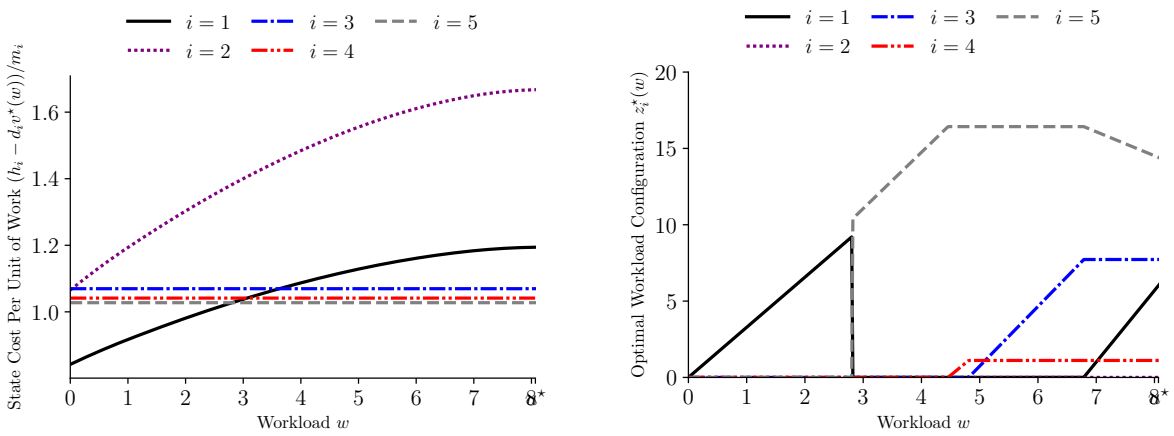
abandonment costs, and waiting costs, respectively. Tables D.10, D.11, D.12, and D.13 report the optimized baseline service rates Δ_0 for the S-OS and D-DO policies.



(a) State cost per unit of work $(h_i - d_i v^\star(w))/m_i$.

(b) Optimal workload configuration $z_i^\star(w)$.

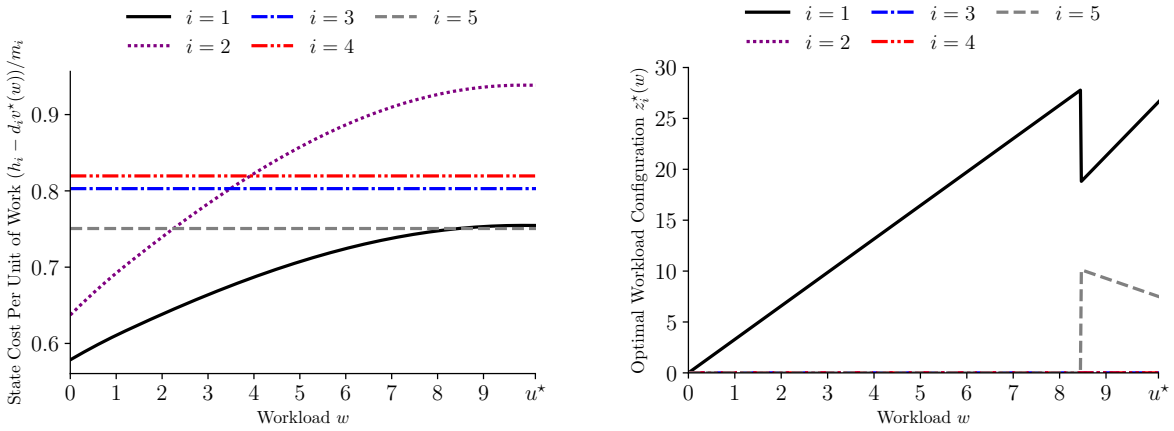
Figure D.3: Effective state cost per unit of work and optimal workload configuration, as a function of workload, with $\Delta_0 = 0.11$ and $\Delta = 0.0$; all other parameters are as in the base cas.



(a) State cost per unit of work $(h_i - d_i v^\star(w))/m_i$.

(b) Optimal workload configuration $z_i^\star(w)$.

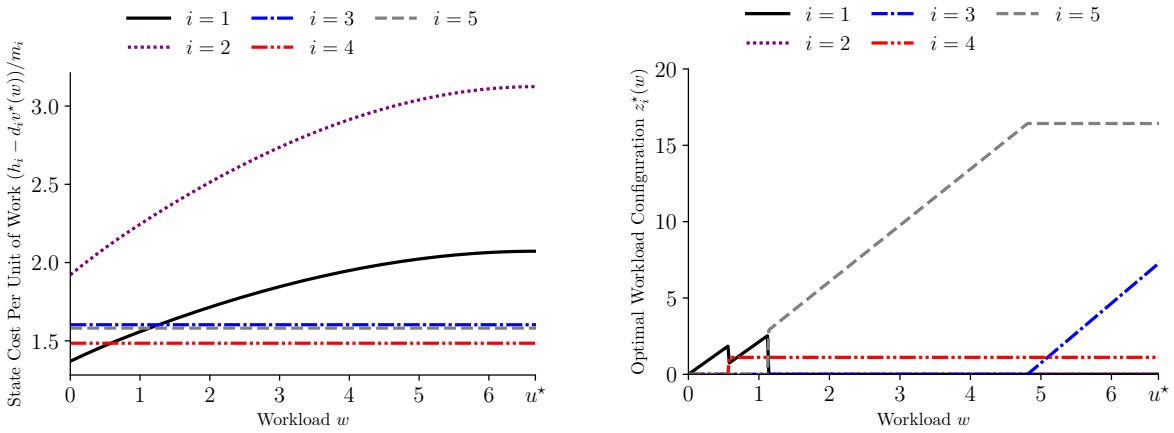
Figure D.4: Effective state cost per unit of work and optimal workload configuration, as a function of workload, with $\Delta_0 = 0.03$ and $\Delta = 0.32$; all other parameters are as in the base case.



(a) State cost per unit of work $(h_i - d_i v^\star(w))/m_i$.

(b) Optimal workload configuration $z_i^\star(w)$.

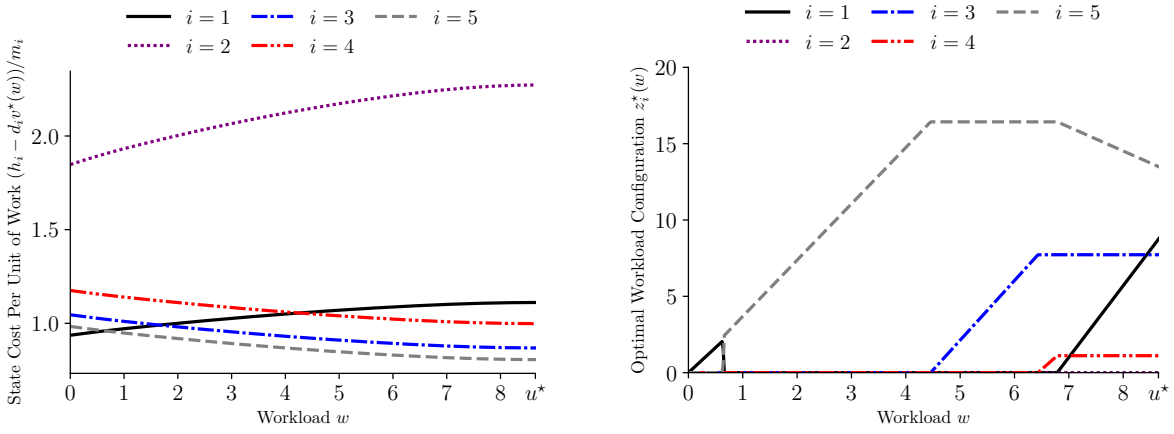
Figure D.5: Effective state cost per unit of work and optimal workload configuration, as a function of workload, with the sensitivity of service outcomes and transitions to waiting times halved; all other parameters are as in the base case.



(a) State cost per unit of work $(h_i - d_i v^\star(w))/m_i$.

(b) Optimal workload configuration $z_i^\star(w)$.

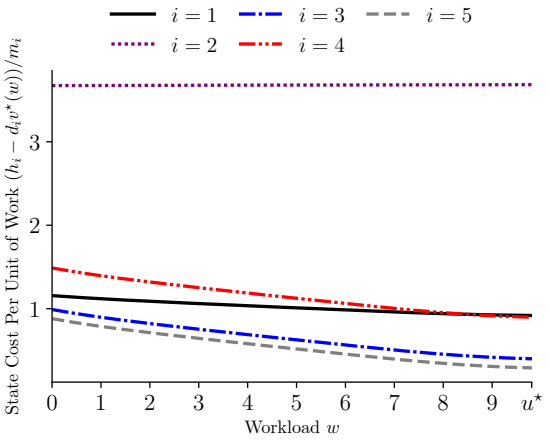
Figure D.6: Effective state cost per unit of work and optimal workload configuration, as a function of workload, with the sensitivity of service outcomes and transitions to waiting times doubled; other parameters are as in the base case.



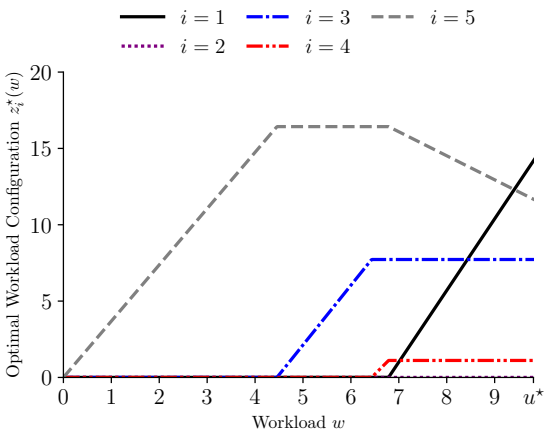
(a) State cost per unit of work $(h_i - d_i v^\star(w))/m_i$.

(b) Optimal workload configuration $z_i^\star(w)$.

Figure D.7: Effective state cost per unit of work and optimal workload configuration, as a function of workload, with abandonment rate $\ell_i = 0.03$ for $i \in \mathcal{I}$; all other parameters are as in the base case.

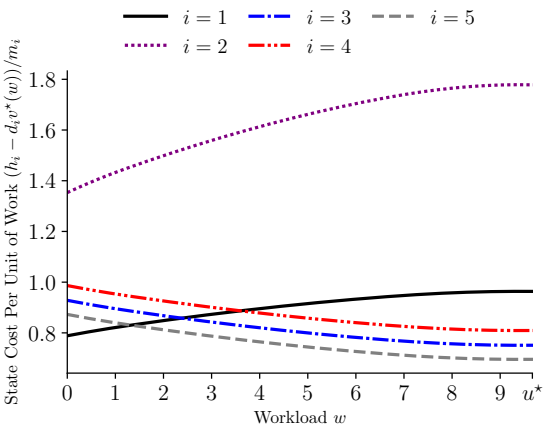


(a) State cost per unit of work $(h_i - d_i v^\star(w))/m_i$.

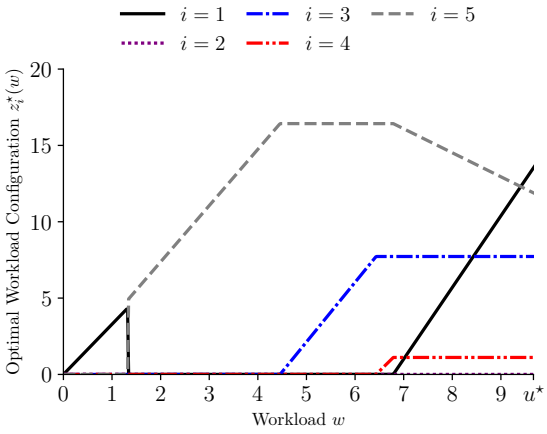


(b) Optimal workload configuration $z_i^\star(w)$.

Figure D.8: Effective state cost per unit of work and optimal workload configuration, as a function of workload, with abandonment rate $\ell_i = 0.1$ for $i \in \mathcal{I}$; all other parameters are as in the base case.

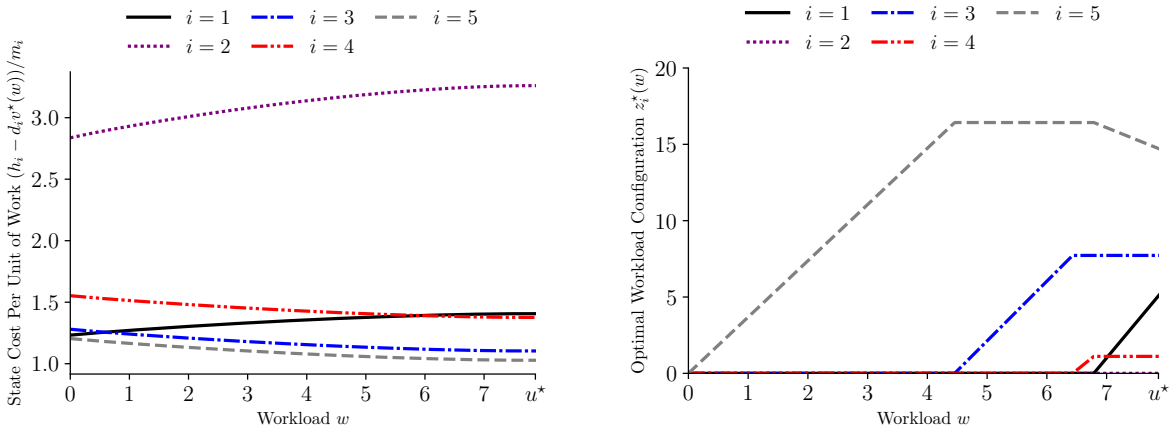


(a) State cost per unit of work $(h_i - d_i v^\star(w))/m_i$.



(b) Optimal workload configuration $z_i^\star(w)$.

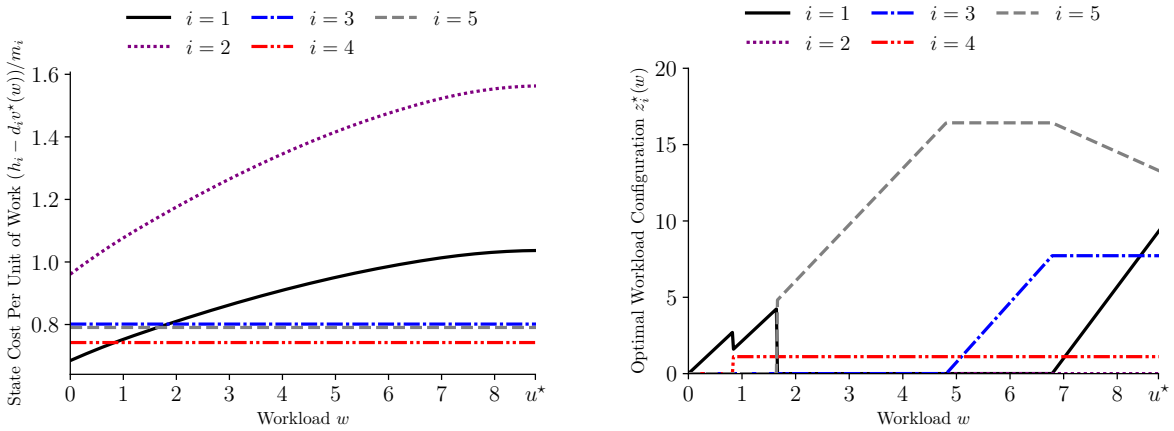
Figure D.9: Effective state cost per unit of work and optimal workload configuration, as a function of workload. The abandonment rates are $\ell_i = 0.03$ for $i \in \mathcal{I}$ and abandonment costs are halved; other parameters are as in the base case.



(a) State cost per unit of work $(h_i - d_i v^\star(w))/m_i$.

(b) Optimal workload configuration $z_i^\star(w)$.

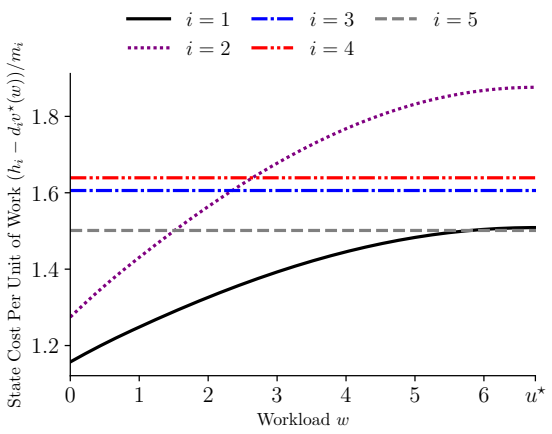
Figure D.10: Effective state cost per unit of work and optimal workload configuration. The Abandonment rates are $\ell_i = 0.03$ for $i \in \mathcal{I}$ and abandonment costs are doubled; other parameters are as in the base case.



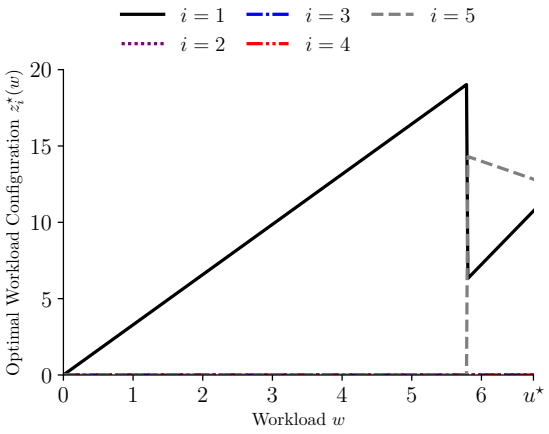
(a) State cost per unit of work $(h_i - d_i v^\star(w))/m_i$.

(b) Optimal workload configuration $z_i^\star(w)$.

Figure D.11: Effective state cost per unit of work and optimal workload configuration, as a function of workload, with waiting costs halved; all other parameters are as in the base case.



(a) State cost per unit of work $(h_i - d_i v^\star(w))/m_i$.



(b) Optimal workload configuration $z_i^\star(w)$.

Figure D.12: Effective state cost per unit of work and optimal workload configuration, as a function of workload, with waiting cost doubled; all other parameters are as in the base case.

Table D.4: Comparative evaluation of the long-run average cost of the various policies for three different deadlines (δ_i for $i \in \mathcal{I}$). Deadlines of all classes are the same.

Deadline	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
1	Static Sch., Nominal Static Rates (S-NS)	66.50 ± 0.14	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	38.48 ± 0.05	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	66.50 ± 0.14	46.24 ± 0.12	37.79 ± 0.07	34.83 ± 0.06	34.29 ± 0.06
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	136.87 ± 0.23	114.19 ± 0.19	96.98 ± 0.17	84.25 ± 0.17	75.45 ± 0.14
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	31.69 ± 0.06	28.79 ± 0.05	26.68 ± 0.05	25.37 ± 0.05	24.86 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	29.64 ± 0.05	27.77 ± 0.05	26.26 ± 0.05	25.28 ± 0.05	24.86 ± 0.05
2	Static Sch., Nominal Static Rates (S-NS)	182.98 ± 0.44	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	38.45 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	182.98 ± 0.44	82.46 ± 0.35	45.01 ± 0.15	35.24 ± 0.08	32.82 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	274.77 ± 0.27	240.75 ± 0.29	207.00 ± 0.29	172.27 ± 0.29	138.63 ± 0.31
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	49.61 ± 0.11	39.87± 0.11	32.03 ± 0.08	27.44 ± 0.07	25.47 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	32.60 ± 0.06	29.67 ± 0.06	27.40 ± 0.05	25.92 ± 0.05	25.22 ± 0.05
3	Static Sch., Nominal Static Rates (S-NS)	321.58 ± 0.94	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	38.58 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	321.58 ± 0.94	128.33 ± 0.71	51.78 ± 0.24	35.84 ± 0.10	32.58 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	414.01 ± 0.29	391.01 ± 0.28	361.40 ± 0.27	334.88 ± 0.28	305.80 ± 0.28
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	59.78 ± 0.15	46.22 ± 0.15	33.94 ± 0.11	27.71 ± 0.08	25.58 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	33.00 ± 0.05	29.85 ± 0.06	27.52 ± 0.05	25.92 ± 0.05	25.18 ± 0.05

Notes. The reported values are sample averages and their associated 95% confidence intervals.

Table D.5: Comparative evaluation of long-run average costs, varying only the sensitivity of service outcomes and transitions to waiting times (c_{1j} and c_{2j} for $j \in \mathcal{J}$).

Sensitivity to Waiting Times	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
Halved	Static Sch., Nominal Static Rates (S-NS)	95.70 ± 0.35	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	30.22 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	95.70 ± 0.35	43.02 ± 0.20	28.93 ± 0.07	25.71 ± 0.06	25.31 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	149.04 ± 0.26	113.89 ± 0.30	82.69 ± 0.22	59.19 ± 0.18	45.73 ± 0.16
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	53.33 ± 0.11	42.07 ± 0.14	28.24 ± 0.15	20.61 ± 0.08	18.78 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	26.95 ± 0.06	23.27 ± 0.05	20.61 ± 0.05	19.19 ± 0.04	18.73 ± 0.05
Base Case	Static Sch., Nominal Static Rates (S-NS)	182.98 ± 0.44	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	38.45 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	182.98 ± 0.44	82.46 ± 0.35	45.01 ± 0.15	35.24 ± 0.08	32.82 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	274.77 ± 0.27	240.75 ± 0.29	207.00 ± 0.29	172.27 ± 0.29	138.63 ± 0.31
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	49.61 ± 0.11	39.87± 0.11	32.03 ± 0.08	27.44 ± 0.07	25.47 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	32.60 ± 0.06	29.67 ± 0.06	27.40 ± 0.05	25.92 ± 0.05	25.22 ± 0.05
Doubled	Static Sch., Nominal Static Rates (S-NS)	363.97 ± 0.87	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	53.95 ± 0.09	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	363.97 ± 0.87	184.14 ± 0.56	94.52 ± 0.34	61.44 ± 0.15	50.29 ± 0.10
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	550.02 ± 0.43	518.02 ± 0.46	480.90 ± 0.35	456.91 ± 0.44	421.77 ± 0.35
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	71.00 ± 0.11	59.17 ± 0.14	48.08 ± 0.12	40.93 ± 0.08	37.90 ± 0.06
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	45.19 ± 0.07	42.00 ± 0.05	39.72 ± 0.06	38.08 ± 0.05	37.16 ± 0.05

Notes. The reported values are sample averages and their associated 95% confidence intervals. We vary c_{11} and keep the ratios c_{1j}/c_{11} and c_{2j}/c_{11} fixed for $j \in \mathcal{J}$.

Table D.6: Comparative evaluation of long-run average costs, varying only the abandonment rate (ℓ_i for $i \in \mathcal{I}$).

Abandonment Rate	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
0.00	Static Sch., Nominal Static Rates (S-NS)	182.98 ± 0.44	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	38.45 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base.(S-DZ)	182.98 ± 0.44	82.46 ± 0.35	45.01 ± 0.15	35.24 ± 0.08	32.82 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	274.77 ± 0.27	240.75 ± 0.29	207.00 ± 0.29	172.27 ± 0.29	138.63 ± 0.31
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	49.61 ± 0.11	39.87± 0.11	32.03 ± 0.08	27.44 ± 0.07	25.47 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	32.60 ± 0.06	29.67 ± 0.06	27.40 ± 0.05	25.92 ± 0.05	25.22 ± 0.05
0.01	Static Sch., Nominal Static Rates (S-NS)	132.28 ± 0.40	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	37.90 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	132.28 ± 0.40	63.93 ± 0.22	40.55 ± 0.11	34.01 ± 0.06	32.44 ± 0.06
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	181.92 ± 0.07	187.49 ± 0.08	180.09 ± 0.11	171.30 ± 0.30	113.17 ± 1.27
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	43.58 ± 0.11	35.17 ± 0.09	29.44 ± 0.08	26.44 ± 0.06	25.39 ± 0.04
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	31.88 ± 0.05	29.14 ± 0.05	27.14 ± 0.05	25.82 ± 0.04	25.23 ± 0.05
0.03	Static Sch., Nominal Static Rates (S-NS)	77.47 ± 0.26	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	37.27 ± 0.05	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	77.47 ± 0.26	46.61 ± 0.14	36.13 ± 0.08	32.83 ± 0.06	32.20 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	174.25 ± 0.11	161.43 ± 0.20	122.97 ± 0.68	46.75 ± 0.40	31.64 ± 0.10
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	34.97 ± 0.07	29.65 ± 0.07	26.77 ± 0.06	25.50 ± 0.05	25.38 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	30.47 ± 0.06	28.08 ± 0.05	26.44 ± 0.05	25.50 ± 0.05	25.29 ± 0.05
0.05	Static Sch., Nominal Static Rates (S-NS)	55.51 ± 0.15	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	37.12 ± 0.05	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	55.51 ± 0.15	40.27 ± 0.09	34.51 ± 0.07	32.55 ± 0.05	32.35 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	31.17 ± 0.06	29.15 ± 0.07	28.27 ± 0.06	28.78 ± 0.06	30.36 ± 0.06
	Dyn. Sch., Dyn. Rates, Zero Base (D-DZ)	29.98 ± 0.05	26.72 ± 0.06	24.88 ± 0.04	24.07 ± 0.05	24.18 ± 0.04
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	28.55 ± 0.05	26.37 ± 0.05	24.88 ± 0.04	24.05 ± 0.05	23.99 ± 0.05
0.07	Static Sch., Nominal Static Rates (S-NS)	46.56 ± 0.10	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	37.25 ± 0.04	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	46.56 ± 0.10	37.80 ± 0.08	34.18 ± 0.05	32.86 ± 0.05	32.93 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	27.33 ± 0.05	29.71 ± 0.07	34.61 ± 0.08	41.36 ± 0.09	48.93 ± 0.08
	Dyn. Sch., Dyn. Rates, Zero Base (D-DZ)	27.34 ± 0.06	25.19 ± 0.05	23.92 ± 0.04	23.48 ± 0.04	23.78 ± 0.04
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	27.14 ± 0.05	25.19 ± 0.05	23.87 ± 0.05	23.31 ± 0.04	23.51 ± 0.04
0.10	Static Sch., Nominal Static Rates (S-NS)	41.93 ± 0.08	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	37.91 ± 0.05	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	41.93 ± 0.08	37.04 ± 0.06	34.75 ± 0.05	34.04 ± 0.05	34.47 ± 0.04
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	25.56 ± 0.04	25.87 ± 0.05	28.35 ± 0.05	33.04 ± 0.07	39.26 ± 0.07
	Dyn. Sch., Dyn. Rates, Zero Base (D-DZ)	25.64 ± 0.04	24.05 ± 0.04	23.15 ± 0.05	22.89 ± 0.04	23.36 ± 0.04
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	25.48 ± 0.05	23.79 ± 0.04	22.78 ± 0.04	22.53 ± 0.04	23.10 ± 0.04

Notes. We assume the abandonment rates of all classes are the same. The reported values are sample averages and their associated 95% confidence intervals.

Table D.7: Comparative evaluation of long-run average costs, varying only the abandonment costs (π_i for $i \in \mathcal{I}$) at the abandonment rate of $\ell_i = 0.03$ for $i \in \mathcal{I}$.

Ratio of Abandonment Cost to Rejection Cost	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
0.25	Static Sch., Nominal Static Rates (S-NS)	66.47 ± 0.24	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	33.95 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	66.47 ± 0.24	40.70 ± 0.13	31.92 ± 0.08	29.40 ± 0.06	29.00 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	154.44 ± 0.10	142.53 ± 0.16	114.94 ± 0.64	54.28 ± 0.38	36.47 ± 0.12
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	31.66 ± 0.08	26.93 ± 0.07	24.40 ± 0.05	23.27 ± 0.05	23.19 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	27.89 ± 0.05	25.67 ± 0.05	24.07 ± 0.04	23.27 ± 0.05	23.19 ± 0.05
0.5	Static Sch., Nominal Static Rates (S-NS)	77.47 ± 0.26	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	37.27 ± 0.05	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	77.47 ± 0.26	46.61 ± 0.14	36.13 ± 0.08	32.83 ± 0.06	32.20 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	174.25 ± 0.11	161.43 ± 0.20	122.97 ± 0.68	46.75 ± 0.40	31.64 ± 0.10
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	34.97 ± 0.07	29.65 ± 0.07	26.77 ± 0.06	25.50 ± 0.05	25.38 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	30.47 ± 0.06	28.08 ± 0.05	26.44 ± 0.05	25.50 ± 0.05	25.29 ± 0.05
0.75	Static Sch., Nominal Static Rates (S-NS)	88.32 ± 0.27	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	40.50 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	88.32 ± 0.27	52.64 ± 0.15	40.23 ± 0.10	36.23 ± 0.05	35.20 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	39.20 ± 0.08	34.68 ± 0.08	31.56 ± 0.07	30.12 ± 0.06	30.30 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	38.23 ± 0.08	32.24 ± 0.07	28.90 ± 0.06	27.44 ± 0.05	27.12 ± 0.04
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	32.69 ± 0.05	30.18 ± 0.05	28.43 ± 0.05	27.36 ± 0.05	27.03 ± 0.05
1.0	Static Sch., Nominal Static Rates (S-NS)	99.19 ± 0.33	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	43.56 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	99.19 ± 0.33	58.62 ± 0.18	44.27 ± 0.09	39.53 ± 0.07	38.25 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	42.58 ± 0.09	37.36 ± 0.09	33.56 ± 0.07	31.58 ± 0.06	31.16 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	41.46 ± 0.10	35.05 ± 0.08	31.34 ± 0.06	29.57 ± 0.05	29.10 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	34.93 ± 0.05	32.45 ± 0.05	30.63 ± 0.05	29.47 ± 0.05	29.05 ± 0.05
1.25	Static Sch., Nominal Static Rates (S-NS)	109.87 ± 0.35	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	46.49 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	109.87 ± 0.35	64.67 ± 0.16	48.25 ± 0.10	42.85 ± 0.06	41.21 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	45.64 ± 0.09	39.86 ± 0.09	35.58 ± 0.07	33.34 ± 0.05	32.69 ± 0.06
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	44.49 ± 0.10	37.73 ± 0.07	33.77 ± 0.07	31.74 ± 0.05	31.20 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	37.13 ± 0.05	34.69 ± 0.06	32.85 ± 0.05	31.69 ± 0.05	31.19 ± 0.05
1.5	Static Sch., Nominal Static Rates (S-NS)	120.74 ± 0.40	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	49.29 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	120.74 ± 0.40	70.48 ± 0.21	52.33 ± 0.11	46.08 ± 0.08	44.06 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	48.00 ± 0.10	40.57 ± 0.09	35.90 ± 0.07	33.56 ± 0.05	32.84 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	47.93 ± 0.11	40.57 ± 0.07	36.10 ± 0.06	34.01 ± 0.05	33.32 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	39.33 ± 0.06	36.91 ± 0.06	35.04 ± 0.05	33.82 ± 0.05	33.31 ± 0.05
2.0	Static Sch., Nominal Static Rates (S-NS)	142.55 ± 0.44	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	54.72 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	142.55 ± 0.44	82.25 ± 0.21	60.30 ± 0.12	52.41 ± 0.07	49.86 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	54.98 ± 0.11	46.10 ± 0.08	40.58 ± 0.08	37.75 ± 0.06	36.67 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	53.96 ± 0.11	45.77 ± 0.09	40.80 ± 0.07	38.35 ± 0.05	37.52 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	43.54 ± 0.06	41.20 ± 0.06	39.39 ± 0.05	38.14 ± 0.05	37.49 ± 0.05

Notes. The abandonment costs are assumed to be the same. The reported values are sample averages and their associated 95% confidence intervals.

Table D.8: Comparative evaluation of long-run average costs, varying only the waiting costs (c_{3j} for $j \in \mathcal{J}$).

Waiting Costs	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
Zero	Static Sch., Nominal Static Rates (S-NS)	138.94 ± 0.47	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	30.57 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	138.94 ± 0.47	58.54 ± 0.26	32.06 ± 0.10	26.12 ± 0.06	25.09 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	253.78 ± 0.15	252.99 ± 0.17	233.50 ± 0.19	225.48 ± 0.19	229.34 ± 0.22
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	28.96 ± 0.08	23.79 ± 0.08	20.32 ± 0.06	18.80 ± 0.05	18.63 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	24.10 ± 0.05	21.44 ± 0.06	19.60 ± 0.06	18.67 ± 0.05	18.63 ± 0.05
Halved	Static Sch., Nominal Static Rates (S-NS)	158.64 ± 0.45	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	34.51 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	158.64 ± 0.45	69.44 ± 0.31	38.11 ± 0.13	30.42 ± 0.07	28.81 ± 0.06
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	266.85 ± 0.24	233.66 ± 0.26	201.13 ± 0.23	171.44 ± 0.25	145.80 ± 0.21
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	37.67 ± 0.08	30.67 ± 0.08	25.87 ± 0.07	23.54 ± 0.06	22.87 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	29.07 ± 0.06	26.07 ± 0.06	24.07 ± 0.05	22.99 ± 0.05	22.80 ± 0.05
Base Case	Static Sch., Nominal Static Rates (S-NS)	182.98 ± 0.44	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	38.45 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	182.98 ± 0.44	82.46 ± 0.35	45.01 ± 0.15	35.24 ± 0.08	32.82 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	274.77 ± 0.27	240.75 ± 0.29	207.00 ± 0.29	172.27 ± 0.29	138.63 ± 0.31
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	49.61 ± 0.11	39.87± 0.11	32.03 ± 0.08	27.44 ± 0.07	25.47 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	32.60 ± 0.06	29.67 ± 0.06	27.40 ± 0.05	25.92 ± 0.05	25.22 ± 0.05
Doubled	Static Sch., Nominal Static Rates (S-NS)	250.76 ± 0.65	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	46.45 ± 0.07	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	250.76 ± 0.65	117.17 ± 0.45	62.11 ± 0.18	46.04 ± 0.10	41.50 ± 0.07
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	296.93 ± 0.29	262.67 ± 0.25	228.48 ± 0.24	189.39 ± 0.31	145.76 ± 0.32
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	82.29 ± 0.12	68.01 ± 0.16	54.25 ± 0.15	40.33 ± 0.13	33.63 ± 0.08
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	40.93 ± 0.07	37.82 ± 0.05	34.63 ± 0.05	32.76 ± 0.05	31.66 ± 0.05

Table D.9: Optimized baseline service rates (Δ_0) the S-OS and D-DO policies in Table D.4.

Deadline	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
1	Static Sch., Opt. Static Rates (S-OS)	0.10	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.06	0.05	0.03	0.01	0.00
2	Static Sch., Opt. Static Rates (S-OS)	0.13	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.11	0.09	0.08	0.05	0.03
3	Static Sch., Opt. Static Rates (S-OS)	0.14	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.12	0.09	0.08	0.05	0.03

Table D.10: Optimized baseline service rates (Δ_0) of the S-OS and D-DO policies in Table D.5.

Sensitivity to Waiting Times	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
Halved	Static Sch., Opt. Static Rates (S-OS)	0.10	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.11	0.09	0.06	0.04	0.01
Base Case	Static Sch., Opt. Static Rates (S-OS)	0.13	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.11	0.09	0.08	0.05	0.03
Doubled	Static Sch., Opt. Static Rates (S-OS)	0.16	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.14	0.11	0.10	0.07	0.04

Table D.11: Optimized baseline service rates (Δ_0) of the S-OS and D-DO policies in Table D.6.

Aband-onment Rate	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
0.00	Static Sch., Opt. Static Rates (S-OS)	0.13	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.11	0.09	0.08	0.05	0.03
0.01	Static Sch., Opt. Static Rates (S-OS)	0.12	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.10	0.08	0.06	0.04	0.02
0.03	Static Sch., Opt. Static Rates (S-OS)	0.10	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.06	0.04	0.03	0.01	-0.01
0.05	Static Sch., Opt. Static Rates (S-OS)	0.09	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.04	0.02	0.00	-0.01	-0.03
0.07	Static Sch., Opt. Static Rates (S-OS)	0.05	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.01	0.00	-0.01	-0.02	-0.04
0.10	Static Sch., Opt. Static Rates (S-OS)	0.05	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	-0.02	-0.02	-0.04	-0.05	-0.07

Table D.12: Optimized baseline service rates (Δ_0) of the S-OS and D-DO policies in Table D.7.

Ratio of Abandonment Cost to Rejection Cost	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
0.25	Static Sch., Opt. Static Rates (S-OS)	0.09	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.06	0.04	0.03	0.00	-0.02
0.5	Static Sch., Opt. Static Rates (S-OS)	0.10	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.06	0.04	0.03	0.00	-0.02
0.75	Static Sch., Opt. Static Rates (S-OS)	0.11	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.07	0.05	0.03	0.01	-0.02
1.0	Static Sch., Opt. Static Rates (S-OS)	0.11	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.08	0.05	0.03	0.01	-0.01
1.25	Static Sch., Opt. Static Rates (S-OS)	0.12	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.08	0.06	0.04	0.01	-0.01
1.5	Static Sch., Opt. Static Rates (S-OS)	0.13	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.08	0.06	0.04	0.02	-0.01
1.75	Static Sch., Opt. Static Rates (S-OS)	0.13	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.09	0.06	0.04	0.02	-0.01
2.0	Static Sch., Opt. Static Rates (S-OS)	0.14	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.09	0.07	0.04	0.02	-0.01

Table D.13: Optimized baseline service rates (Δ_0) of the S-OS and D-DO policies in Table D.8.

Waiting Costs	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
Zero	Static Sch., Opt. Static Rates (S-OS)	0.11	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.07	0.05	0.04	0.02	0.00
Halved	Static Sch., Opt. Static Rates (S-OS)	0.12	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.09	0.07	0.05	0.04	0.03
Base Case	Static Sch., Opt. Static Rates (S-OS)	0.13	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.11	0.09	0.08	0.05	0.03
Doubled	Static Sch., Opt. Static Rates (S-OS)	0.15	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.16	0.13	0.11	0.07	0.04

Table D.14: Comparative evaluation of the long-run average cost of the various policies for three different deadlines (δ_i for $i \in \mathcal{I}$) in the discrete non-preemptive setting of Section 8.5.

Deadline	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
1	Static Sch., Nominal Static Rates (S-NS)	175.54 ± 0.16	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	45.32 ± 0.06	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	175.54 ± 0.16	143.26 ± 0.17	110.82 ± 0.16	79.10 ± 0.19	54.05 ± 0.11
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	247.13 ± 0.23	208.71 ± 0.25	171.54 ± 0.22	139.36 ± 0.22	113.28 ± 0.19
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	37.22 ± 0.08	29.33 ± 0.06	23.82 ± 0.06	20.86 ± 0.05	19.93 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	26.78 ± 0.04	23.77 ± 0.04	21.63 ± 0.04	20.37 ± 0.04	19.93 ± 0.05
2	Static Sch., Nominal Static Rates (S-NS)	283.47 ± 0.15	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	48.09 ± 0.05	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base.(S-DZ)	283.47 ± 0.15	255.80 ± 0.13	229.06 ± 0.16	201.08 ± 0.15	161.44 ± 0.30
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	420.09 ± 0.27	383.02 ± 0.25	346.10 ± 0.24	309.41 ± 0.24	271.67 ± 0.24
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	58.48 ± 0.13	41.79 ± 0.12	28.28 ± 0.09	22.09 ± 0.07	20.24 ± 0.04
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	28.43 ± 0.07	24.63 ± 0.05	22.06 ± 0.04	20.55 ± 0.04	20.07 ± 0.04
3	Static Sch., Nominal Static Rates (S-NS)	366.34 ± 0.16	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	48.48 ± 0.09	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	366.34 ± 0.16	344.03 ± 0.13	322.06 ± 0.14	296.47 ± 0.17	255.39 ± 0.23
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	542.90 ± 0.21	526.92 ± 0.23	497.59 ± 0.24	474.67 ± 0.25	442.34 ± 0.26
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	78.54 ± 0.13	52.26 ± 0.21	29.72 ± 0.12	22.25 ± 0.07	20.23 ± 0.04
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	28.60 ± 0.05	24.81 ± 0.04	22.10 ± 0.04	20.61 ± 0.04	20.04 ± 0.04

Notes. Deadlines of all classes are the same. The reported values are sample averages and their associated 95% confidence intervals.

Table D.15: Comparative evaluation of long-run average costs, varying only the sensitivity of service outcomes and transitions to waiting times (c_{1j} and c_{2j} for $j \in \mathcal{J}$) in the discrete non-preemptive setting of Section 8.5.

Sensitivity to Waiting Times	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
Halved	Static Sch., Nominal Static Rates (S-NS)	186.92 ± 0.16	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	38.17 ± 0.09	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	186.92 ± 0.16	157.24 ± 0.14	128.00 ± 0.16	94.13 ± 0.21	49.60 ± 0.24
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	257.88 ± 0.25	217.81 ± 0.24	178.02 ± 0.23	135.15 ± 0.26	92.69 ± 0.23
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	69.52 ± 0.12	51.62 ± 0.18	22.79 ± 0.16	15.11 ± 0.05	14.86 ± 0.04
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	22.65 ± 0.06	18.52 ± 0.04	15.89 ± 0.05	14.77 ± 0.04	14.86 ± 0.04
Base Case	Static Sch., Nominal Static Rates (S-NS)	283.47 ± 0.15	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	48.09 ± 0.05	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base.(S-DZ)	283.47 ± 0.15	255.80 ± 0.13	229.06 ± 0.16	201.08 ± 0.15	161.44 ± 0.30
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	420.09 ± 0.27	383.02 ± 0.25	346.10 ± 0.24	309.41 ± 0.24	271.67 ± 0.24
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	58.48 ± 0.13	41.79 ± 0.12	28.28 ± 0.09	22.09 ± 0.07	20.24 ± 0.04
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	28.43 ± 0.07	24.63 ± 0.05	22.06 ± 0.04	20.55 ± 0.04	20.07 ± 0.04
Doubled	Static Sch., Nominal Static Rates (S-NS)	449.40 ± 0.18	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	61.11 ± 0.07	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	449.40 ± 0.18	418.73 ± 0.17	385.12 ± 0.20	341.87 ± 0.22	273.18 ± 0.37
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	766.16 ± 0.39	741.17 ± 0.45	699.29 ± 0.41	699.09 ± 0.63	652.47 ± 0.53
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	85.97 ± 0.15	64.80 ± 0.17	44.55 ± 0.13	35.21 ± 0.07	32.08 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	39.23 ± 0.05	36.08 ± 0.05	33.75 ± 0.05	32.08 ± 0.05	31.39 ± 0.04

Notes. We vary c_{11} and keep the ratios c_{1j}/c_{11} and c_{2j}/c_{11} fixed for $j \in \mathcal{J}$. The reported values are sample averages and their associated 95% confidence intervals.

Table D.16: Comparative evaluation of long-run average costs, varying only the abandonment rate (ℓ_i for $i \in \mathcal{I}$) in the discrete non-preemptive setting of Section 8.5.

Abandonment Rate	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
0.00	Static Sch., Nominal Static Rates (S-NS)	283.47 ± 0.15	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	48.09 ± 0.05	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base.(S-DZ)	283.47 ± 0.15	255.80 ± 0.13	229.06 ± 0.16	201.08 ± 0.15	161.44 ± 0.30
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	420.09 ± 0.27	383.02 ± 0.25	346.10 ± 0.24	309.41 ± 0.24	271.67 ± 0.24
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	58.48 ± 0.13	41.79 ± 0.12	28.28 ± 0.09	22.09 ± 0.07	20.24 ± 0.04
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	28.43 ± 0.07	24.63 ± 0.05	22.06 ± 0.04	20.55 ± 0.04	20.07 ± 0.04
0.01	Static Sch., Nominal Static Rates (S-NS)	272.42 ± 0.15	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	47.76 ± 0.09	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	272.42 ± 0.15	245.00 ± 0.14	217.98 ± 0.15	187.82 ± 0.21	135.78 ± 0.32
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	248.45 ± 0.10	263.07 ± 0.10	248.88 ± 0.10	243.41 ± 0.12	203.39 ± 0.22
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	49.83 ± 0.12	35.87 ± 0.11	26.45 ± 0.07	22.39 ± 0.05	21.01 ± 0.05
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	28.40 ± 0.05	25.01 ± 0.05	22.61 ± 0.04	21.12 ± 0.04	20.63 ± 0.04
0.03	Static Sch., Nominal Static Rates (S-NS)	251.35 ± 0.14	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	47.55 ± 0.09	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	251.35 ± 0.14	224.15 ± 0.13	195.15 ± 0.17	154.13 ± 0.25	83.37 ± 0.35
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	241.43 ± 0.14	222.04 ± 0.14	192.59 ± 0.22	39.31 ± 0.12	27.95 ± 0.06
	Dyn. Sch., Dyn. Rates, Zero Base. (D-DZ)	37.47 ± 0.09	29.25 ± 0.06	25.79 ± 0.05	25.02 ± 0.04	25.54 ± 0.04
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	28.76 ± 0.05	26.64 ± 0.05	25.79 ± 0.05	25.02 ± 0.04	25.54 ± 0.04
0.05	Static Sch., Nominal Static Rates (S-NS)	231.43 ± 0.17	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	47.54 ± 0.05	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	231.43 ± 0.17	202.89 ± 0.15	167.05 ± 0.21	109.54 ± 0.31	60.67 ± 0.17
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	37.23 ± 0.07	33.25 ± 0.06	32.16 ± 0.06	33.17 ± 0.05	35.10 ± 0.04
	Dyn. Sch., Dyn. Rates, Zero Base (D-DZ)	30.05 ± 0.08	25.06 ± 0.06	22.85 ± 0.05	22.41 ± 0.04	22.95 ± 0.04
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	26.44 ± 0.04	24.12 ± 0.04	22.79 ± 0.05	22.41 ± 0.04	22.95 ± 0.04
0.07	Static Sch., Nominal Static Rates (S-NS)	211.54 ± 0.16	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	48.00 ± 0.05	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	211.54 ± 0.16	178.75 ± 0.19	131.70 ± 0.25	79.26 ± 0.24	56.43 ± 0.11
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	26.50 ± 0.05	25.79 ± 0.05	26.69 ± 0.05	28.89 ± 0.05	31.58 ± 0.04
	Dyn. Sch., Dyn. Rates, Zero Base (D-DZ)	26.46 ± 0.06	23.44 ± 0.05	22.18 ± 0.04	22.14 ± 0.04	22.96 ± 0.04
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	25.40 ± 0.04	23.24 ± 0.04	22.18 ± 0.04	22.14 ± 0.04	22.96 ± 0.04
0.10	Static Sch., Nominal Static Rates (S-NS)	178.07 ± 0.20	–	–	–	–
	Static Sch., Opt. Static Rates (S-OS)	49.34 ± 0.05	–	–	–	–
	Static Sch., Dyn. Rates, Zero Base. (S-DZ)	178.07 ± 0.20	134.34 ± 0.23	91.01 ± 0.21	67.95 ± 0.11	60.70 ± 0.07
	Dyn. Sch., Dyn. Rates, Zero Base., Ign. Wait	24.09 ± 0.04	24.39 ± 0.05	26.20 ± 0.05	29.04 ± 0.05	32.48 ± 0.05
	Dyn. Sch., Dyn. Rates, Zero Base (D-DZ)	24.10 ± 0.04	22.41 ± 0.04	21.70 ± 0.04	21.94 ± 0.04	23.07 ± 0.04
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	24.10 ± 0.04	22.41 ± 0.04	21.70 ± 0.04	21.94 ± 0.04	23.07 ± 0.04

Notes. We assume the abandonment rates of all classes are the same. The reported values are sample averages and their associated 95% confidence intervals.

Table D.17: Optimized baseline service rates (Δ_0) of the S-OS and D-DO policies in Table D.14.

Deadline	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
1	Static Sch., Opt. Static Rates (S-OS)	0.20	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.10	0.08	0.06	0.02	0.00
2	Static Sch., Opt. Static Rates (S-OS)	0.24	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.10	0.08	0.06	0.04	0.02
3	Static Sch., Opt. Static Rates (S-OS)	0.24	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.12	0.10	0.06	0.06	0.02

Table D.18: Optimized baseline service rates (Δ_0) of the S-OS and D-DO policies in Table D.15.

Sensitivity to Waiting Times	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
Halved	Static Sch., Opt. Static Rates (S-OS)	0.18	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.10	0.08	0.04	0.02	0.00
Base Case	Static Sch., Opt. Static Rates (S-OS)	0.24	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.10	0.08	0.06	0.04	0.02
Doubled	Static Sch., Opt. Static Rates (S-OS)	0.28	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.14	0.10	0.08	0.06	0.04

Table D.19: Optimized baseline service rates (Δ_0) of the S-OS and D-DO policies in Table D.16.

Aband-onment Rate	Policy	Maximum Service Rate Variation				
		0.00	0.08	0.16	0.24	0.32
0.00	Static Sch., Opt. Static Rates (S-OS)	0.24	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.10	0.08	0.06	0.04	0.02
0.01	Static Sch., Opt. Static Rates (S-OS)	0.22	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.10	0.08	0.06	0.04	0.04
0.03	Static Sch., Opt. Static Rates (S-OS)	0.18	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.06	0.04	0.02	0.00	0.00
0.05	Static Sch., Opt. Static Rates (S-OS)	0.20	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.06	0.04	0.02	0.00	0.00
0.07	Static Sch., Opt. Static Rates (S-OS)	0.18	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.04	0.02	0.00	0.00	0.00
0.10	Static Sch., Opt. Static Rates (S-OS)	0.16	–	–	–	–
	Dyn. Sch., Dyn. Rates, Opt. Base. (D-DO)	0.00	0.00	0.00	0.00	0.00