Optimal Dynamic Control of Queueing Networks: Emergency Departments, the W Service Network, and Supply Chains under Disruptions

by

Soroush Saghafian

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in The University of Michigan
2012

Doctoral Committee:

Associate Professor Mark P. Van Oyen, Chair
Professor Wallace J. Hopp
Professor Xiuli Chao
Associate Professor Hyun-Soo Ahn

To my parents and sister with deepest gratitude
for their sincere love and continuous support.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my dissertation committee chair Professor Mark Van Oyen. I am especially grateful for his incredible continuous support and guidance. He introduced me to an exciting research area which ultimately resulted in this dissertation. He has also been more than a thesis advisor for me, providing me with advice on a broader level.

I have also been greatly privileged to work closely under the supervision of Professor Wallace Hopp. I owe a debt of gratitude to him. His guidance and support have been unparalleled: he has been always a source of inspiration for me, generous in his time, and truly supportive, especially when needed most. He has contributed enormously in shaping and advancing my thinking. By being inspirational, he has always lifted my motivations when I have encountered barriers.

I would also like to thank Professor Xiuli Chao for his encouragement and advice. I have been flattered to have the opportunity of learning various materials in the field of Queueing Theory and Stochastic Processes from him. I would like to also express my appreciation to my other dissertation committee member Professor Hyun-Soo Ahn for his comments and support. I have also benefited from the advice and guidance of Professor Demosthenis Teneketzis who has been always supportive of my work and has taught me much of my knowledge in the general area of Stochastic Control.

I would like to also express my sincere gratitude to my parents for their support and encouragement. Also, I would like to express my deep appreciation to my girlfriend, Elyse Jennings, for her incredible support during the time I have been working on this dissertation. Finally, I am deeply thankful to Siamak Davarani and my other friends for their help and support.

TABLE OF CONTENTS

DEDICATION ii

LIST OF FIGURES vii

LIST OF TABLES x

CHAPTERS

1 Introduction 1
- 1.1 Overview 1
- 1.2 Motivation and Research Objectives 2
  - 1.2.1 Organization of the Dissertation 5

2 Emergency Department: Patient Streaming 6
- 2.1 Introduction 6
- 2.2 Literature Survey 9
- 2.3 Modeling Flows and Performance in the ED 12
- 2.4 Phase 1 Implications of Streaming and Pooling 16
- 2.5 Phase 2 Implications of Streaming and Pooling 22
- 2.6 A Simulation-Based Comparison of Streaming and Pooling . 26
  - 2.6.1 ED Flow Design: Pooling, Physical Streaming, or Virtual Streaming? 31
  - 2.6.2 Sensitivity Analyses: Where to Implement Virtual Streaming? 36
- 2.7 Conclusion 42
- 2.8 Appendix A: Proofs. 44
- 2.9 Appendix B: Computations Under Imperfect Classification . 60
- 2.10 Appendix C: Further Descriptions of the Simulation Framework and Assumptions. 69

3 Emergency Department: Complexity-Based Triage 74
- 3.1 Introduction 74

3.2 Literature Review . . . 79
3.2.1 Operations Research/Management Studies . . . 79
3.2.2 Medical Studies . . . 81
3.3 Modeling the ED . . . 83
3.4 Phase 1: A Simplified Single-Stage ED Model . . . 85
3.4.1 Urgency-Based Triage - Phase 1 . . . 86
3.4.2 Complexity-Based Triage - Phase 1 . . . 88
3.5 Phase 2: A Multi-Stage ED Model . . . 91
3.6 A Realistic Simulation Analysis of Complexity-Based Triage . 96
3.6.1 Performance of Complexity-Based Triage . . . 101
3.6.2 How to Define Complex Patients? . . . 102
3.6.3 The Effect of ED Resource Levels . . . 103
3.6.4 The Effect of Misclassification . . . 105
3.6.5 Complexity-Based Streaming . . . 105
3.7 Conclusion . . . 107
3.8 Appendix (Proofs) . . . 109

4 Dynamic Control in the W Service Network and Beyond . . . 118
4.1 Introduction . . . 118
4.2 Literature Survey . . . 122
4.3 General Characteristics . . . 123
4.3.1 The Model . . . 124
4.3.2 Formulation of the Markov Decision Process . . . 125
4.3.3 Stability . . . 127
4.4 The "W" Structure . . . 130
4.4.1 The "W" Structure: An Efficient System Design . . . 130
4.4.2 Dynamic Control of Servers in the "W" Structure . . 132
4.4.3 An Efficient Heuristic Policy: Largest Expected Workload Cost (LEWC) . . . 135
4.4.4 Computational Results . . . 137
4.5 Conclusion . . . 145
4.6 Appendix A: Proofs . . . 147
4.7 Appendix B: Structure of the Optimal Policy . . . 176
4.8 Appendix C: Test Suites for Numerical Comparisons . . . 179
4.8.1 PART I: Comparisons of Various Structures (Figures 4.3 and 4.4) . . . 180
4.8.2 Part II: Comparisons of LEWC with Other Well-Known Policies . . . 180

5 Supply Chain Disruption Risk Management: Newsvendor Analysis with Recourse . . . 188
5.1 Introduction . . . 188

5.2 Related Literature . . . 191
5.3 Model and Notation . . . 193
5.4 Analyses with Recourse (Two-Stage Setting) . . . 195
5.4.1 Single–Product Special Case . . . 198
5.4.2 Two–Product Case . . . 201
5.4.3 Recourse Analysis with Offshore Unreliable Suppliers 208
5.5 Benchmark Analyses: No Recourse . . . 211
5.5.1 Benchmark Setting: The Value of the Secondary Flexible Backup Supplier . . . 212
5.5.2 Benchmark Setting: The Value of Disruption Risk Information . . . 214
5.6 The Value of Recourse, Flexibility, and Information . . . 217
5.6.1 The Value of Recourse . . . 217
5.6.2 The Value of Flexibility . . . 219
5.6.3 The Value of Disruption Risk Information . . . 220
5.7 Summary of Findings and Conclusion . . . 221
5.8 Appendix A: Proofs . . . 224
5.9 Appendix B: Parameter Settings . . . 251
5.10 Appendix C: Optimal Capacity Reservation Levels in Study 1 253
5.11 Appendix D: Further Results on the Two-Product Setting with Recourse . . . 259

6 Supply Chain Disruption Risk Management: Dynamic Analyses . . . 262
6.1 Introduction . . . 262
6.2 The Literature . . . 265
6.3 The Model . . . 267
6.4 Analyses . . . 271
6.4.1 Full Information on Disruption Risk Levels . . . 272
6.4.2 No Information on Disruption Risk Levels . . . 278
6.4.3 The Value of Flexibility: A Single Flexible Backup Supplier or Dedicated Ones? . . . 283
6.5 Summary and Conclusion . . . 288
6.6 Appendix A: Proofs . . . 290
6.7 Appendix B: Parameter Settings . . . 295

BIBLIOGRAPHY . . . 299

LIST OF FIGURES

Figure

2.1 The general flow of patients in an ED. 13

2.2 An example of the optimal strategy with three admissible policies (PA, PD, S) and deterministic service times for which streaming is almost never optimal. 19

2.3 Sensitivity of policies to changes in α. Streaming is more robust to changes in patient mix than are the pooling policies. 21

2.4 When the level of uncertainty in the percentage of A patients (measured by σ_α^2) increases, streaming becomes the optimal policy for an increasingly wide range of β values. 22

2.5 Multi-stage ED service: (W(1): (initial) wait, T(1):(initial) treatment, FW: final wait, E: exit). 25

2.6 Class dependent arrival rates to the ED for an average day (obtained from a year of data in UMED). 27

2.7 Virtual streaming significantly outperforms pooling and improved pooling. The reason is that VS dramatically decreases LOS for 3D patients with only a minor increase in TTFT of 2A patients. (Results for an ED with 8 physicians and 60 beds, a 20% misclassification error rate, and a weight for TTFT of A patients of $\beta = 0.50$). 32

2.8 The benefit of implementing PNPO sequencing rule. ED's with a higher physician utilization or with a higher maximum number of patients allowed per physician benefit more from PNPO. 37

2.9 Sensitivity of virtual streaming and pooling designs. Lower weight on TTFT of A patients (β) or misclassification probability favors virtual streaming over pooling. 38

2.10 Sensitivity of virtual streaming (VS) and pooling (P) designs with respect to mean μ_α (left) and variance σ_α^2 (right) of the percentage of A patients. VS dominates P in the shaded region. 39

2.11 The effect of average patient test time (MRI, CT Scan, etc.) on the relative performance of two virtual streaming and pooling configurations. As test time for A patients increases (left) or decreases for D patients (right), virtual streaming becomes more attractive compared to pooling. 40
2.12 The effect of treatment times on the relative performance of two virtual streaming and pooling configurations. As treatment time for A patients increases (left) or decreases for D patients (right), virtual streaming becomes more attractive compared to pooling. 40
2.13 The effect of the average boarding time on the performance of two virtual streaming and pooling configurations. ED's with longer boarding of A's benefit more from virtual streaming. 41
2.14 The effect of average physician utilization on the attractiveness of virtual streaming. ED's with higher average physician utilization benefit more. 42
2.15 ED patient flow design strategy based on key environmental characteristics of the ED. 43
2.16 Expected performance of policies for a clearing system with $n = 20$, $\mu_A = 80(mins)$, $\mu_D = 45(mins)$, and symmetric misclassification error between A and D patients. Streaming is more robust to misclassification errors than pooling. 67
2.17 Cumulative number of class based physician-patient interactions . . . 70
3.1 Left: Current practice of triage (Emergency Severity Index (ESI) algorithm version 4); Right: Proposed complexity-based triage system (RU: Resuscitation Unit, FT: Fast Track, NS: Non-urgent Simple, NC: Non-urgent Complex, US: Urgent Simple, UC: Urgent Complex). . . . 78
3.2 General flow of patients in the main ED. 84
3.3 Benefit of complexity-based triage over urgency-based triage with practical misclassification rates ($\gamma_{\mathbf{U}} = \gamma_{\mathbf{N}} = \mathbf{10}\%, \gamma_{\mathbf{S}} = \gamma_{\mathbf{C}} = \mathbf{17}\%$) reported in the literature 91
3.4 Patient flow after a patient is moved to an examination room/bed (Phase 2 sequencing). 94
3.5 Class dependent arrival rates to the ED for an average day (obtained from a year of data in UMED). 97
3.6 Cumulative number of class-based physician-patient interactions for complex patients (those requiring more than one interaction). 99
3.7 Performance of complexity-based triage when defining complex patients to be those having more than one, more than two, and more than three physician-patient interactions. 103

3.8 The effect of resources (beds and physicians) on the benefit of complexity-based triage over the current practice of urgency-based triage [Left: the effect of beds (4 physicians); Right: the effect of physicians (22 beds)]. 104
3.9 The effect of complexity misclassification error rates on the benefit of a complexity-based triage (compared to an urgency-based only) [Left: symmetric misclassification; Right: asymmetric misclassification].) . . 104
3.10 Performance of different patient flow designs compared to the current practice (Urgency-Based Pooling). 107
4.1 An example of a small customer support center (the "W" structure). 121
4.2 A general parallel queueing system with server disruptions and arbitrary flexibility structure. 124
4.3 Various possible structures with $|\mathcal{N}_c| = 3$ and $|\mathcal{N}_s| = 2$. 132
4.4 Comparison of possible structures under four suites of parameters using the optimal policies. 133
4.5 Performance of $c\mu$, LQ, $Gc\mu$, and LEWC relative to the optimal policy. 139
4.6 Sensitivity of $c\mu$, LQ, $Gc\mu$, and LEWC to variations in disruption risks. 145
4.7 Illustration of the optimal policy in examples 1 and 2. 178
4.8 Illustration of the optimal policy in example 3. 179
5.1 The two-echelon supply chain under consideration. 194
5.2 The Value of a Flexible Backup Supplier ($IP_{(F)}\%$) in Settings 1-3 for Different Values of Error in the Firm's Reliability Belief (ϵ^1, ϵ^2) 257
5.3 The Value of Information ($IP_{(I)}\%$) for Different Values of Firm's Reliability Belief Error (ϵ^1, ϵ^2). 258
6.1 The general model under consideration. 269
6.2 The Pooling Effect under Full Information ($\beta = 0.9, c^1 = c^2 = 2, r_1 = r_2 = 3.5, p_1 = p_2 = 2.5, h_1 = 1, h_2 = 1.2$). 288

LIST OF TABLES

Table

2.1 Different patient flow designs under consideration and the notation implemented. 30

2.2 Performance (in hrs) of the proposed streaming design (VS/AD+ESI/PNPO) and current pooling practice (P/ESI/SIRO) under four metrics as well as the associated LWBS (%). For the streaming design the physician and bed split have been optimized at phys. =(3, 5) and beds=(22,38) for the A and D sides, respectively. 35

4.1 Comparison of policies based on the combinations of disruption and the system congestion using the percentage optimality gaps. 143

4.2 Comparison of policies based on the holding cost settings (Level of cost asymmetry among different classes: (A) Zero, (B) Low, (C) Moderate, (D) High). 143

4.3 Various Suites of Parameter Settings for Comparisons of Alternate Structures. 181

4.4 Various service rate combinations and arrival rates in the test suites. 183

4.5 Workload distributions on skills used in Table 4.4 (S: Server; C: Class of Customer). 184

4.6 Settings of disruption rates, repair rates, and system congestion factor (ρ) in addition to the corresponding system utilization. 184

4.7 Holding cost settings and the corresponding highest and second highest $c\mu$ rankings among the queues (SF: Highest=Shared, Second Highest=Fixed; FS: Highest=Fixed, Second Highest=Shared, FF: First and Second Highest=Fixed). 185

4.8 Service and arrival rates considered for the results on the effect of server disruption risk (Fig. 4.6). 186

4.9 Holding cost setting considered for the results on the effect of server disruption risk (Fig. 4.6). 186

4.10 Disruption and repair rate setting considered for the results on the effect of server disruption risk (Fig. 4.6). 187

5.1 The Value of Recourse and the Difference in Investment in the Flexible Backup Capacity with and without Recourse. 219
5.2 The Value of Flexibility and the Reduction in Investment in the Backup Capacity due to Flexibility. 220
5.3 The Value of Disruption Risk Information Under Recourse. 221
5.4 Suite of Parameter Settings in Studies 4 and 5. 252
5.5 Suite of Parameter Settings in Study 1. 252
6.1 Numerical Study 1 (No-Information.) [ϵ^j: firm's reliability perception error of unreliable supplier j]. 283
6.2 The Value of Supply Flexibility under No-Information. 284
6.3 The Value of Supply Flexibility with Full-Information and the Pure Value of Information (without Flexibility). 285
6.4 Suite of Parameter Settings in Study 1. 296
6.5 Suite of Parameter Settings in Study 2. 297
6.6 Suite of t.p.m. Settings. 298

CHAPTER 1

Introduction

1.1 Overview

The first and the main unifying theme of this dissertation is the use of innovative techniques in the design and control of queueing systems, where design and control decisions significantly influence system performance. Special attention is given to real-world applications in (a) the design and control of patient flow in the hospital emergency departments, (b) the design and control of call centers, and (c) the design and control of supply chains. With respect to application (a), we show how better patient prioritization and enhanced patient flow designs allows emergency departments to significantly improve their performance. In application (b), we investigate how effective control and design strategies that make use of (partial) flexibility of servers can be implemented to achieve high performance regimes. In application (c), we study how supply chains can boost their performance using better control and design strategies that specifically take into account disruption risks.

A second unifying theme is the use of better information in making superior design and control decisions. For instance, in application (a), we demonstrate that enhanced triaged systems that collect additional up-front information regarding patient ultimate disposition and/or medical complexity can yield better control and design

mechanisms. Regarding application (b), we show how information about server disruptions (e.g., absenteeism) can be used to improve the performance of the system by using this information in a more sophisticated dynamic control policy. In application (c), we study how collecting disruption risk information (through monitoring unreliable suppliers) can enhance the overall performance.

A third unifying theme is the use of flexibility in the design and control. For instance, we propose to create bed and physician flexibility in emergency departments (application (a)). We show how this flexibility will yield higher resource utilizations and thereby improve the performance. In application (b), we study how much server flexibility is required to achieve a good performance. Finally, with respect to application (c), we show how the use of flexibility in the supply system can create superior supply chains that can respond to disruption risks.

1.2 Motivation and Research Objectives

Many systems in both service and manufacturing sectors can be modeled and analyzed as queueing networks. In such systems control and design can be an important issue that may significantly affect the performance. An example is hospital Emergency Departments (ED's). Between 1996 and 2006, annual visits to ED's in the U.S. increased by 32% (from 90.3 million to 119.2 million), while the number of hospital ED's decreased from 4,019 to 3,833 ([114]). This trend has elevated ED overcrowding to crisis levels in many U.S. hospitals. Similar trends have intensified pressure on ED's around the world. The consequences of ED overcrowding can be tragic. For example, in 2006, 49-year-old Beatrice Vance arrived at the busy ED of Vista Medical Center East in Waukegan, IL, complaining of nausea, shortness of breath, and chest pain. Triaged and sent to the ED waiting room, Mrs. Vance waited there for two hours without further attention. When she was finally called, she failed to respond and was found dead of an acute myocardial infarction ([134]).

The most direct way to alleviate crowding and improve responsiveness is by adding resources. But, since this is also the most expensive approach, it is generally not the preferred option. Analysis in control of queueing systems, however, can be implemented to improve the performance of ED's without adding expensive capacity. These analyses, however, need to be tailored to match practice and to consider detailed specifications in ED's that may deviate from traditional assumptions in queueing networks (e.g., as will be discussed, physicians do not take more than seven patients at the time in practice, or the objective function in the ED is different for different patient types).

Chapters 2 and 3 devise tailored analyses for the ED's. Specifically, Chapter 2 shows how a new patient flow design that separates the service stream for those predicted to be admitted to the hospital and those predicted to be discharged home can improve ED metrics such as Length of Stay (LOS), Time to First Treatment (TTFT), and percentage of patients Left Without Being Seen (LWBS). It is also shown how stochastic control analysis can be used to determine effective patient prioritization for (a) bringing patients from the waiting area into examination rooms, and (b) decide which patient to visit next while they are in examination rooms. In addition to tailored queueing control analysis, a high-fidelity simulation calibrated with a year of data is developed to gain further insights.

Chapter 3 builds upon the previous chapter and suggests a new triage system that revolutionizes the current practice of triage in ED's by collecting information regarding the medical complexity of patients (in addition to their urgency). Given the new triage system, it is found that modified versions of the $c\mu$ rule that take into account patient misclassifications can be effectively implemented to increase both patient safety (i.e., reduce risk of adverse events) and operational efficiency (i.e., reduce length of stay).

Managment of service center operations is another application where control and design plays a significant role. The use of partially flexible servers or cross-trained agents makes the design and control a challenging issue in call centers. Considering the

use of such partially flexible resources, the literature can be divided to the following three main themes: (1) *System Design* of specific paradigms for creating flexibility to maximize an objective, (2) *Server Scheduling and Control* policies to reap the benefits of flexibility, and (3) *Performance Analysis* of specific systems and policies. Chapter 4 contributes to all three themes, especially the second theme. In this chapter, a special attention is given to a structure with two servers and three customer classes forming a "W." In this structure, servers are trained to serve a shared task in addition to their fixed (specialized) task. Similar to some other structures studied in the literature, the "W" design is highly efficient; it requires only a small amount of cross-training, but performs almost as well as a fully cross-trained system. We show that (even allowing disruptions) a version of the $c\mu$ rule, which prioritizes serving the "fixed task before the shared," is optimal under some conditions. However, our results show that, in general, the optimal policy is complex, and not well-structured for implementation in practice. Hence, a powerful and yet very simple policy to control the servers in parallel queueing systems is proposed. This heuristic (which can also be implemented in any parallel queueing system) effectively combines the intuition underlying two widely used policies: (1) the load balancing objective in serving the Longest Queue (LQ), and (2) the greedy cost minimization emphasis of the $c\mu$ rule. Our proposed control policy, termed as "LEWC," defines a simple and intuitive measure of workload costs and assigns each server to the queue with the Largest Expected Workload Cost (LEWC) among its skill set. Our extensive numerical tests show that the proposed policy performs well in comparison with four key policies: optimal, LQ, $c\mu$, and generalized $c\mu$ ($Gc\mu$). We also provide a proof of stability of the LEWC, LQ, and $Gc\mu$ policies.

Another important application where design and control plays a major role is in supply chains. Consider the 2007 disruption in Boeing's 787 Dreamliner's supply chain. Advanced Integration Technology (AIT) fell months behind building parts needed to assemble the plane, thereby wreaking havoc upon Boeing's 787 inflexible

supply chain. Boeing itself expected to take a multi-billion dollar cash hit in 2008 from paying penalties to airline customers and to keep its suppliers afloat in the wake of the serious cash flow disruption, according to Greising and Johnsson (2007). Another example is the disruption in the Toyota supply chain on Feb. 01, 1997. A fire at the Aisin Seiki Co. destroyed most of the capacity to manufacture P-valves. Because of the Aisin's ability to produce parts at low cost, Toyota had come to rely on Aisin for this product (Sheffi, 2007) resulting in a poor supply chain design with limited control options. Chapters 5 and 6 consider the design and control of supply chains under disruption risks and provide new insights that can yield resilience in supply chains.

1.2.1 Organization of the Dissertation

The dissertation is presented in a multiple manuscript format. The results in Chapters 2, 3, 4, 5, and 6 have appeared as individual research papers [121], [120], [124], [122], and [123], respectively. The organization of the dissertation is as follows. Chapters 2 and 3 consider the application of emergency departments. Chapter 4 considers the application of the call centers, Chapters 5 and 6 consider the design and control of supply chains under disruption risks.

CHAPTER 2

Emergency Department: Patient Streaming

2.1 Introduction

Between 1996 and 2006, annual visits to Emergency Departments (ED's) in the U.S. increased by 32% (from 90.3 million to 119.2 million), while the number of hospital ED's decreased from 4,019 to 3,833 ([114]). This trend has elevated ED overcrowding to crisis levels in many U.S. hospitals. Similar trends have intensified pressure on ED's around the world.

The consequences of ED overcrowding can be tragic. For example, in 2006, 49-year-old Beatrice Vance arrived at the busy ED of Vista Medical Center East in Waukegan, IL, complaining of nausea, shortness of breath, and chest pain. Triaged and sent to the ED waiting room, Mrs. Vance waited there for two hours without further attention. When she was finally called, she failed to respond and was found dead of an acute myocardial infarction ([134]).

Other, less tragic but still important, consequences of ED overcrowding include patient "elopement" (i.e., leaving without being seen), ambulance diversions, and treatment delays ([69]). The ED overcrowding situation has become so dire that the American College of Emergency Physicians (ACEP) in its 2006 report gave a failing mark to emergency care in 41 of 50 states in the U.S, and a D- nationally for access

to care (see [4]). Some experts believe that the recent healthcare bill will exacerbate the already serious overcrowding problem in U.S. ED's ([135]).

This situation has prompted researchers to investigate a variety of methods for alleviating ED overcrowding, including: (1) personnel staffing, (2) hospital bed access control, (3) non-urgent and low acuity patient referrals, (4) ambulance diversion, (5) destination control, and (6) improved resource utilization ([69]).

The most direct way to alleviate crowding and improve responsiveness is by adding resources. But, since this is also the most expensive approach, it is generally not the preferred option. Recognizing this, [116] concluded, *"the debate is no longer about the level of resources our EDs deserve, but rather about how to ensure that ED resources are directed to those who need them - the patients in the waiting room.* To achieve this, some practitioners have recently suggested streaming patients based on their likelihood of being admitted to the hospital. In one pioneering effort, Flinders Medical Center in Australia implemented a system in which ED patients and resources are divided into two streams: one for those likely to be discharged (hereafter "Discharge" or "D" patients) and one for those likely to be admitted to the hospital (hereafter "Admit" or "A" patients) ([90], [18]). They reported a 48 minute reduction in average time spent by the patients in the ED. While Flinders is an Australian hospital, the fundamental operational principles governing ED flow design are very similar in developed countries. However, since Flinders represented a single uncontrolled experiment in a specific environment in which other changes (e.g., lean initiatives) were implemented along with the streaming system, it is impossible to infer that their results are purely due to streaming and/or that they will translate to other ED's. Nevertheless, motivated by positive reports from Flinders, other hospitals such as Bendigo Health ED ([91]) have begun implementing similar strategies.

While streaming patients based on the likelihood of being admitted to the hospital is new, patient streaming is not. By the 1980s most ED's (although not Flinders) had adopted separate "fast tracks" for patients with minor injuries ([159]). In the

1990s, many ED's also established "observation units" for patients requiring lengthy diagnosis. But, as [159] noted, *"these innovations were the tip of the iceberg, and performance-driven emergency departments have been experimenting with models that segment patients into streams for more efficient health care delivery."*

For clarity, we will use the term "streaming" to refer specifically to the newly proposed policy that separates patients (and resources) into different streams according to anticipated disposition (A or D). We label the conventional policy that treats both types of patients together (with pooled resources) as "pooling". It is well known from the Operations Management literature that pooling offers efficiency benefits resulting from improved resource utilization. This means that in order for streaming to be effective, it must offer advantages that offset its inherent "anti-pooling" disadvantage. The Flinders results suggest that this may be possible. But since their results could be due to (a) specific conditions (e.g., high percentage of admits, the fact that they did not yet have a separate stream (fast track) for low acuity patients, etc.), (b) other changes (e.g., lean), or (c) a Hawthorne effect halo, we cannot say without a careful analysis.

In this chapter, we use a combination of analytical and simulation models to perform a systematic study of the attractiveness of streaming. Specifically, we address the following questions:

1. *Whether streaming (or a variation on it) can improve ED performance?*
2. *Where (i.e., in what hospital environments) is streaming (or an effective variation on it) most attractive?*
3. *How should Admit/Discharge information be implemented for maximum effectiveness?*

The remainder of this chapter is organized as follows. Section 3.2 summarizes previous research relevant to the above questions. Section 6.3 describes ED flows and

performance metrics in order to construct models with which to understand them. Section 2.4 considers a simple clearing model with a single stage service process, in which patients can be classified (A or D) without error. This analysis provides insight into the relative effectiveness of streaming and pooling with respect to sequencing patients into the examination rooms. While this suggests that sequencing alone is not enough to overcome the anti-pooling disadvantage of streaming, it also indicates that streaming is more robust to patient mix variation and classification errors than is pooling, which can lead to streaming outperforming pooling in real-world settings. In Section 2.5, we consider another analytic clearing model, with perfect patient classification but with multi-stage service processes, in order to understand the impact of patient sequencing within the exam rooms (i.e., the order in which physicians visit the patients assigned to them) on the streaming versus pooling comparison. We find that prioritizing downstream (i.e., near service completion) D patients and upstream (i.e., recent arrivals) A type patients enhances the advantage of streaming over pooling. In Section 2.6, we use a simulation model of a realistic ED environment that includes dynamic patient arrivals, multi-stage service processes, and patient misclassification error to test the conjectures made from our analytic models. Taken together, our results suggest that, implemented properly in the right environment, streaming can significantly improve overall ED performance by substantially reducing wait times for D patients at the expense of only a modest increase in wait times for A patients. We conclude in Section 6.5 with a summary of our overall insights about whether, where, and how streaming can be a potentially attractive strategy for improving ED responsiveness.

2.2 Literature Survey

There are two main streams of research related to the work of this chapter: (1) Empirical studies of the ED overcrowding problem (published in medical journals),

and (2) General queueing systems research (published in operations research journals) that deal with pooling and/or customer sequencing. We highlight key contributions from each of these below.

For an excellent survey of empirical studies of ED overcrowding see [69]. Some of these studies have examined the nature and extent of the problem. For example, [98] showed that there is a strong correlation between ED length of stay and inpatient length of stay and concluded that *"strategies to reduce the length of stay in the ED may significantly reduce healthcare expenditures and patient morbidity"*. The Centers for Disease Control and Prevention (CDC) estimated that 379,000 deaths occurred in U.S. ED's in 2000 ([103]). Other studies have found that long waiting times are linked to patient mortality as well as elevated risks of errors and adverse events (e.g., [143], [51], and [148]). One such study estimated that long waiting times and high occupancies cause 13 deaths per year in one Australian ED ([117]). Thus, reducing waiting times is a means for promoting higher levels of patient safety. Because admit patients typically include the most critical cases that need more rapid attention, some researchers have focused specifically on studying mortality among admit patients. For instance, [136] associated a combined measure of hospital and ED crowding (which causes long waiting times) with an increased risk of mortality among admitted patients.

Other studies have evaluated the factors that influence overcrowding. [106] evaluated different internal factors that affect patient flow and concluded that ED overcrowding is driven by both external pressure and internal factors such as how flow across the ED is measured. [125] studied the effect of low complexity ED patients on the waiting times of other patients and concluded that the impact is negligible. Still other papers have examined the impact of various reorganizations. The papers on the Flinders experiment with streaming ([90], [18]) fall into this category. Another example is [74], which considered a new ED admission process in which ED physicians admit patients directly to the general medical unit after a telephone consultation with

a hospitalist.

A subcategory of empirical research on the ED deals with developing metrics with which to address the issues of ED crowding. [132] provided an overview of the various metrics that have been proposed. We focus on two important measures in our study: Length of Stay (LOS), which measures total time in the ED from arrival to discharge/admit, and Time to First Treatment (TTFT), which measures the time from arrival to the first meaningful interaction with the physician.

Finally, a stream of empirical ED research involves time studies that characterize how caregivers spend their time in the ED, as well as the nature and duration of treatments. Examples of this type of research include [68] and [52]. We will make use of these results to calibrate our models.

A number of researchers have used queueing models to study various aspects of the ED (see, for instance, [30], [54], and [3]). Within the large literature on queueing, studies that consider resource pooling, customer partitioning, or customer sequencing/prioritizing are most relevant to our work. The standard insight from studies of pooling in queueing systems is that when two classes of customers in a queueing system become sufficiently different, pooling becomes ineffective and may even be harmful (see [101], [142], [150]). This suggests that a significant difference in treatment times between A and D type patients may be one way for streaming to overcome the anti-pooling disadvantage. But verifying this requires an extension of known results because in the ED patient misclassification is inevitable, service is a complex process involving several physician-patient interactions, different streams of patients have different performance metrics, and the system has limited buffers (i.e., examination rooms/beds).

A related stream of queueing systems research considers effective ways of partitioning resources (see e.g., [118], [161], [75]). An important observation from these studies is that separating fast and slow customers can protect customers with short processing times from waiting behind customers with long processing times. Note,

however, that the same effect can be achieved by assigning priorities to customers with shorter processing times ([75]). However, for either partitioning or prioritizing to work effectively, we must be able to classify customers with a high level of accuracy. Analyses of priority queueing systems under misclassification errors (which are inevitable in ED's) suggest that these insights may not hold when classification is imperfect (see, e.g., [11]).

One last line of queueing research relevant to our work is the one that studies sequencing. In queueing systems where multiple customers are in the system at the same time (e.g., serial production lines with jobs at different stages of competition or an ED with multiple patients in the exam rooms awaiting physician attention), the server (physician) faces a customer sequencing problem. Related studies of serial systems can be found in [40], [70], and [152], while related studies of parallel queueing systems can be found, for instance, in [9], [124], and the references therein. In particular, [152] proposed a "pick-and-run" policy for servers in a serial system which favors working on the most downstream (old) jobs. We find that a similar policy can help physicians assigned to the D stream to choose their next patient in a manner that reduces average LOS.

2.3 Modeling Flows and Performance in the ED

To develop a modeling framework with which to address the *whether*, *where*, and *how* questions stated above, we must first describe the key characteristics of ED operations. We start by representing the general flow of patients in Figure 2.1. Patients arrive to the ED in a non-stationary, stochastic manner. Upon arrival, patients first go to the triage stage where each patient is assigned an Emergency Severity Index (ESI), usually by a nurse but sometimes by a doctor. ESI is an integer between 1 to 5, where clinical urgency decreases in ESI level. ESI 1 patients (who constitute a small percentage of total patient volume) are subject to high mortality risk if not treated

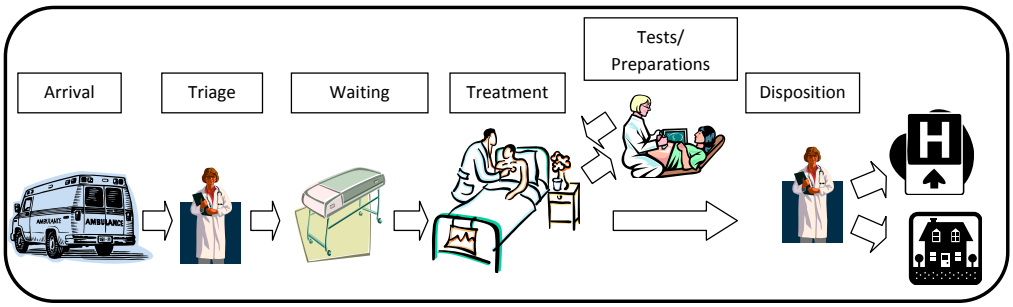


Figure 2.1: The general flow of patients in an ED.

immediately. Hence, they are always given high priority. As such, they are generally tracked separately from the rest of the patients through an "acute care" or "resuscitation" track. In American hospitals, ESI 4 and 5 patients are also often tracked separately through a "fast track" because their treatment needs are relatively simple and straightforward. Hence, in this chapter, we focus on the ESI 2 and 3 patients who make up the bulk (about 80% at the University of Michigan) of the patients in the main ED.

In addition to assigning ESI levels, Flinders Medical Center has reported that, at the time of triage, nurses can predict whether a patient is A or D with roughly 80% accuracy ([90]). Empirical studies in other medical centers have reported similar results (e.g., [66], [94]).

After a patient has been triaged, he/she waits in a waiting area, and is eventually called to an examination room. There he/she goes through one or more phases of interaction (treatment) with the same physician, as shown in Figure 2.1. (While caregivers may be non-physicians (e.g., physician assistants), we use the term "physician" for simplicity.) Each physician-patient interaction (treatment stage) lasts a stochastic amount of time and is followed by testing (MRI, CT scan, etc.) or processing activities (e.g., wound cleaning) by a nurse that do not involve the physician. During testing or processing stages, which are also stochastic in duration, the patient is unavailable to the physician. The final processing stage after the last physician interaction is "disposition," in which the patient is either discharged or admitted to

the hospital by staff based on the physician's final instructions.

Note that a patient is usually assigned to a single physician and so must wait for his/her physician to return for each treatment phase. Also, in most ED's, a patient is assigned to an exam room and holds that room, even when he/she is sent to a test facility, until he/she is disposed (discharged or admitted). Since physicians and exam rooms are limited, both of these resources can be bottlenecks.

The flow of patients in the ED is impacted by two phases of sequencing decisions. *Phase 1 sequencing* decisions determine the order/priority in which patients are initially taken from the waiting area to an examination room. Phase 1 decisions are usually made by a nurse in consideration of ESI levels and patient arrival orders. In theory, it could also make use of A/D predictions. Once patients are in examination rooms, *Phase 2 sequencing* is done to determine the order in which patients are seen. Individual physicians make the Phase 2 sequencing decisions by choosing the patients assigned to them in consideration of ESI levels, patient comfort, time in system, experience, etc. We have observed wide variance in the Phase 2 sequencing logic of individual physicians working within the same ED. Furthermore, physicians tend to limit the number of patients they have at any given time – seven seems to be a typical upper limit.

It is impossible to capture all of the above-mentioned complexities of the ED in a single tractable analytic model. Of course, we can use simulation, but it is difficult to draw clear insights from purely numerical studies. Therefore, to probe the *whether*, *where*, and *how* questions, we will first examine a series of analytic models that represent simplified versions of the ED flow and then test the resulting conclusions under realistic conditions with a high fidelity simulation calibrated with hospital data.

To compare streaming and pooling strategies, we must model the flows under each protocol. In a typical ED, which uses a pooling protocol, patients are not classified into A/D categories and all (ESI 2 and 3) patients are served by a set of pooled/shared resources (exam rooms, physicians, etc.), with priority given to ESI 2 patients. Under

the streaming protocol, resources are divided into two groups: one for the A stream and one for the D stream, and A/D predictions are used to direct patients to the appropriate stream.

To compare the pooling and streaming protocols, we also need a performance criterion. Two commonly used metrics in the ED are Length of Stay (LOS) and Time to First Treatment (TTFT). For D patients, LOS is the key metric because it correlates with both convenience and safety (since a low LOS also guarantees a low TTFT). But for A patients, LOS in the ED is usually a small fraction of their total LOS in the hospital, which on average extends for days beyond their time in the ED. For these patients, safety is of much greater importance than amount of time they spend in the ED rather than in a hospital bed. Since safety is enhanced by starting treatment as soon as possible, TTFT is the most important metric for A patients.

We let α denote the percentage of A patients and define $T_A^\pi(\alpha)$ and $L_D^\pi(\alpha)$ to be the (average) TTFT of A patients and (average) LOS of D patients under policy $\pi \in \mathbf{\Pi}$, respectively, where $\mathbf{\Pi} = \{PA(\text{Pooling with priority to A's}),$
$PD(\text{Pooling with priority to D's}), S(\text{Streaming})\}$ represents the set of admissible policies. More specifically, letting N denote the total number of patients who visit the ED during a sufficiently long period (e.g., a year), we define $T_A^\pi(\alpha) = \mathbb{E}^\pi[\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} T_{A,i}]$ and $L_D^\pi(\alpha) = \mathbb{E}^\pi[\frac{1}{(1-\alpha)N}\sum_{i=1}^{(1-\alpha)N} L_{D,i}]$, where αN $((1-\alpha)N)$ is the number of A's (D's) during the period, $\mathbb{E}^\pi$ denotes expectation with respect to the probability measure defined by policy $\pi \in \Pi$, and $T_{A,i}$ and $L_{D,i}$ are random variables denoting the TTFT and LOS of the ith A and the ith D patient, respectively. Note that we are restricting attention only to pooling and streaming policies, in keeping with the "whether" question raised in the Introduction. We acknowledge that a more complex state-dependent policy might outperform the policies in set $\mathbf{\Pi}$. But how much improvement is possible and whether such policies can be made practical in actual ED settings are open questions. In this chapter, we restrict our attention to the potential for improvement through demonstrably implementable streaming policies.

To construct a single objective function, we let β represent the relative weight placed on the TTFT of A patients and define $f^{\pi}(\alpha, \beta) = \beta\, T_A^{\pi}(\alpha) + (1-\beta) L_D^{\pi}(\alpha)$ as the performance metric under policy $\pi \in \mathbf{\Pi}$. We note that this performance metric can also be derived from a cost perspective. To see this, suppose c_A and c_D represent the per patient cost of increasing the TTTF of A patients and LOS of D patients by one unit of time, respectively. If $\beta = (c_A\,\alpha)/\big(c_A\,\alpha + c_D\,(1-\alpha)\big)$, then $f^{\pi}(\alpha, \beta)$ represents the average cost per patient under policy π. For instance, setting $\beta = \alpha$ implies an objective in which increasing TTTF of A patients and LOS of D patients by one minute is equally costly. We also note that while other metrics are used to evaluate the performance of an ED, most of these are highly correlated with our objective function. For example, the percentage of patients who leave without being seen (LWBS) is commonly tracked in ED's, but studies such as [44] have indicated that the majority of such patients leave the ED because of prolonged waiting times. Hence, improvements in our objective function can be expected to result in reduced LWBS as well. We will examine the impact of streaming on LWBS in Section 2.6.

A closer look at the empirical results reported by Flinders ([90]) indicates that streaming reduced the LOS of D patients but increased TTFT of A patients. Hence, if streaming is attractive, it is because it strikes a better balance between these potentially conflicting objectives. Our combined objective enables us to examine this tradeoff.

2.4 Phase 1 Implications of Streaming and Pooling

Realistic models of ED flow described in the previous section would be too complex for anything other than simulation. So, to get some clear insights into *whether*, *where*, and *how* streaming can outperform pooling, we start with a stylized patient flow model in which (1) all patients are available at the beginning of each day (i.e., static arrivals), (2) there are only two physicians, who work in parallel under the pooling

protocol and are assigned to the A and D streams in the streaming protocol, (3) patient diagnosis/treatment occurs in a merged single service stage, where X_A (X_D) is a random variable with mean μ_A (μ_D) representing the service time of an A (D) patient, (4) A/D classification is perfect (i.e., error free), and (5) to avoid inefficient underutilization, the A(D) physician switches to serve D(A) patients when there is no other A(D) patient is available. Because we model service as a single stage, we eliminate the Phase 2 sequencing decisions. Hence, this model only offers insights into the performance of pooling and streaming via their impact on Phase 1 sequencing.

The above assumptions (most of which will be relaxed in subsequent sections) allow us to represent the ED with a *clearing* queueing model, in which a fixed number (n) of patients is available at the beginning of the day. Because the overall performance of the ED is heavily influenced by performance during periods of overload (which occur during predictably in the mid afternoon), the clearing model approximates ED behavior better than the more conventionally used steady state queueing model.

We start by examining the relative effectiveness of the three policies in the admissible space $\mathbf{\Pi}$ for extreme cases where $\beta = 1$ or 0 (i.e., when the objective function is either merely TTFT for A's or LOS for D's).

Proposition 1 (Extreme Cases) *With $\mathbf{\Pi} = \{PA, PD, S\}$, the following hold for the clearing model (with arbitrary distributions of X_A and X_D) :*

(i) For every $\alpha \in [0,1]$ and every sample path ω, $\operatorname{argmin}_{\pi\in\mathbf{\Pi}} T_A^\pi(\alpha,\omega) = PA$. *That is, if only TTFT of A's matters (i.e., when $\beta = 1$), then pooling with priority to A's is the best policy in $\mathbf{\Pi}$ (in the almost sure sense).*

(ii) For every $\alpha \in [0,1]$ and every sample path ω, $\operatorname{argmin}_{\pi\in\mathbf{\Pi}} L_D^\pi(\alpha,\omega) = PD$. *That is, if only LOS of D's matters (i.e., when $\beta = 0$), then pooling with priority to D's is the best policy in $\mathbf{\Pi}$ (in the almost sure sense).*

This intuitive proposition suggests that streaming is not attractive unless we care about both TTFT for A's and LOS for D's. Therefore, we now analyze the optimal strategy when the objective function is a convex combination of these two metrics. To do this, we first formally define a *strategy* for our problem.

Definition 1 (Strategy) *A* strategy *is a map* $\pi : [0,1] \times [0,1] \to \mathbf{\Pi}$ *that defines the policy* $\pi(\alpha, \beta)$ *for each* α, β*. An optimal strategy is the one that defines an optimal policy* $\pi^*(\alpha, \beta)$ *for every* (α, β)*.*

A useful property that allows us to establish the structure of the optimal strategy is β-convexity which we define in two steps as follows.

Definition 2 (π Region) *For policy* $\pi \in \mathbf{\Pi}$*, the* π *region, denoted by* $\mathcal{A}^\pi$*, is the collection of* (α, β) *for which policy* π *is optimal. That is, the* π *region is* $\mathcal{A}^\pi = \{(\alpha, \beta) \in [0,1] \times [0,1] : \pi^*(\alpha, \beta) = \pi\}$*.*

Definition 3 (β-Convexity) *The optimal strategy* $\pi^* : [0,1] \times [0,1] \to \mathbf{\Pi}$ *is said to be* β-convex *if all the* π *regions (i.e, sets* $\mathcal{A}^\pi$ *(*$\forall \pi \in \mathbf{\Pi}$*)) are convex in* β *for every* $\alpha \in [0,1]$*.*

Lemma 1 (β-Convexity) *The optimal strategy* $\pi^*(\alpha, \beta)$ *is* β*-convex.*

Using the above lemma, we can establish the structure of the optimal strategy.

Proposition 2 (Double Threshold Policy) *For every fixed* $\alpha \in [0,1]$*, there exist double thresholds* $\underline{\beta}(\alpha), \bar{\beta}(\alpha)$ *such that streaming is the best policy in* $\mathbf{\Pi}$ *if, and only if,* $\beta \in [\underline{\beta}(\alpha), \bar{\beta}(\alpha)]$*. If* $\beta < \underline{\beta}(\alpha)$ *then pooling with priority to D's is the best policy in* $\mathbf{\Pi}$*. If* $\beta > \bar{\beta}(\alpha)$ *then pooling with priority to A's is the best policy in* $\mathbf{\Pi}$*.*

Since ED's vary in their percentage of A's (α) and relative weight of TTFT of D's (β), the appeal of streaming depends on the width of the gap between $\underline{\beta}$ and $\bar{\beta}$. Unfortunately, our numerical experiments suggest that this gap is very narrow

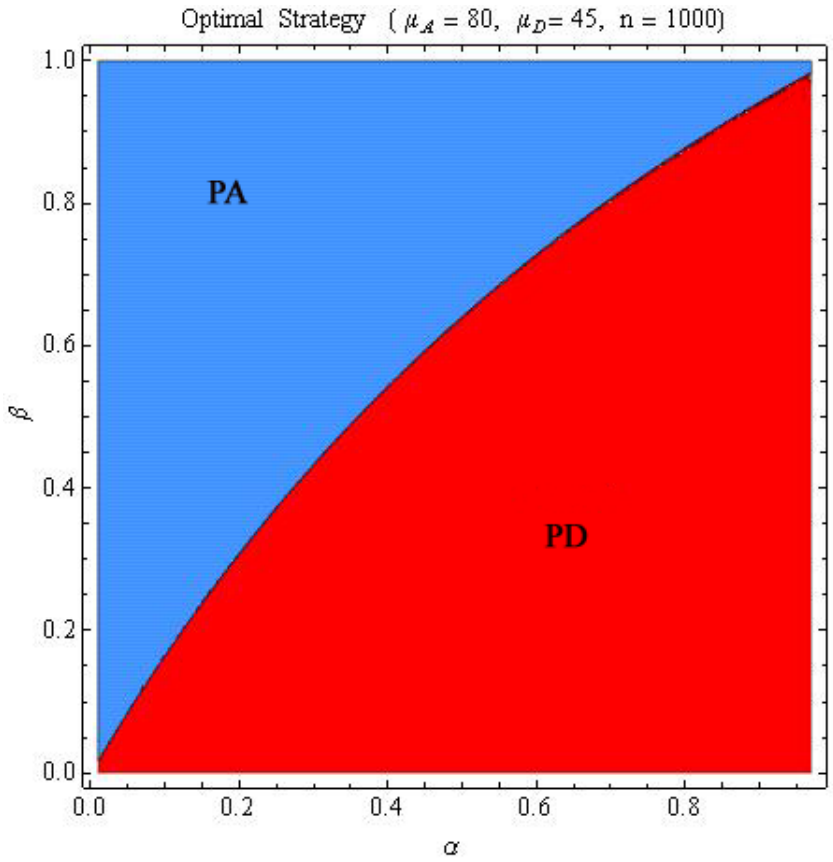


Figure 2.2: An example of the optimal strategy with three admissible policies (PA, PD, S) and deterministic service times for which streaming is almost never optimal.

for the stylized model of this section. Indeed, Figure 2.2 illustrates an example with deterministic service times in which there is no region of optimality for streaming (it can, however, appear with stochastic service times). While the optimality region for streaming can appear when service times are stochastic, it is generally small when α is constant and known. Knowing the exact proportion of A's enables a fixed priority policy to strike an effective balance between the waiting costs of A's and D's.

This is no longer the case under the (highly realistic) assumption that α is uncertain. If the percentage of A patients varies from day to day, then a pooling policy that prioritizes either A or D patients can be quite ineffective. The reason is that we must pick which patient type to prioritize before the mix of A and D patients is known for the day. If we choose the wrong policy for the mix that actually occurs,

performance could be very poor. We illustrate this in Figure 2.3, which plots the optimality gap (i.e., difference between the objective function of a given policy and that of the optimal policy) for the S, PA and PD policies. These results show that while PA is optimal for small α, it is very poor for large α. Conversely, PD is optimal for large α and very poor for small α. In contrast, the streaming policy, S, is almost never optimal but is also never poor. Hence, we can make the following observation.

Observation 1 *Streaming is much more* robust *to changes in patient mix (α) than is pooling.*

The reason is that streaming mimics a dynamic policy with the simplicity of a static rule. By allocating some capacity to both patient types, it never results in a few patients of one type waiting for many patients of the other type.

To examine the impact of uncertainty in α, we assume α is chosen from a family of Beta distributions given by $\text{Beta}\big(f(x), 2f(x)\big)$, where $f(x) = (2-9x)/(27x)$, $x \in (0, 2/9)$. This results in $\mu_\alpha = 1/3$, which approximates the fraction of A's in the University of Michigan Emergency Department (UMED), and $\sigma_\alpha^2 = x$, so we can generate a range of uncertainty of α by varying x. We choose the Beta distribution because (1) it is the most common distribution for a random variable that takes values between zero and one, and it includes the other well-known distribution, the uniform, as a special case, and (2) it seems to well represent our data from UMED. Figure 2.4 uses our analytical model of the ED along with the Beta distribution to illustrate the impact of varying σ_α^2 on the optimal strategy. This figure offers two insights: (1) As noted before, when there is no uncertainty (σ_α^2=0), streaming is not optimal for any value of β. (2) As the level of uncertainty (measured by σ_α^2) increases, streaming becomes optimal for an increasingly broad range of β values.

From Figures 2.3 and 2.4, we can make an important conjecture (which we will test in Section 2.6): streaming is more robust than pooling to variation in patient mix. The intuition behind this robustness result is that a pooling system that completely

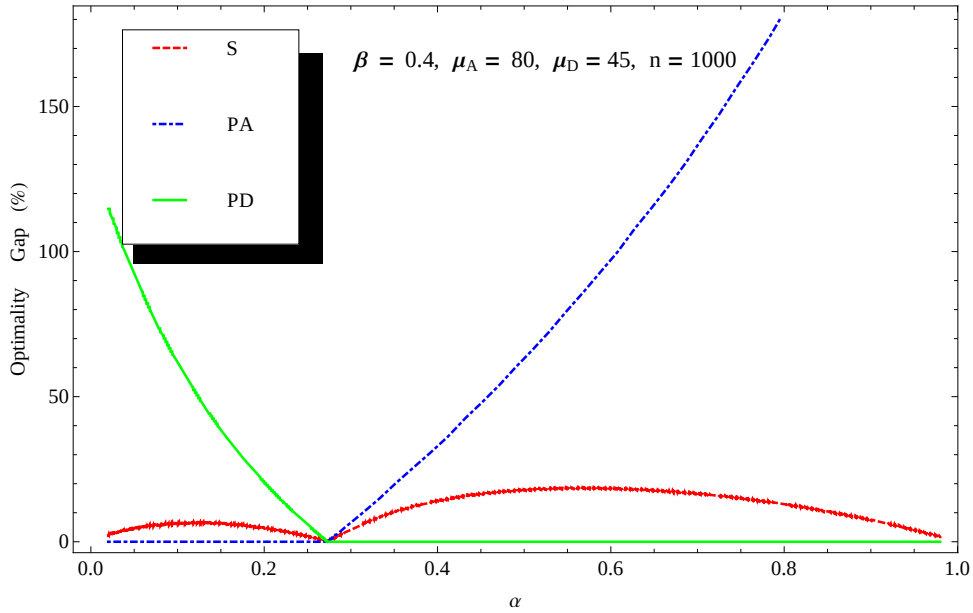


Figure 2.3: Sensitivity of policies to changes in α. Streaming is more robust to changes in patient mix than are the pooling policies.

prioritizes one type of patients can sequence them far from the optimal order (e.g., putting D patients at the end of the line on a day in which they should have been at the beginning of the line). In contrast, a streaming system always gives some priority to both types of patients by "reserving" some capacity for each type. While the proportion of capacity assigned to A and D patients may not be optimal on any given day (depending on the mix of patients), the fact that the two streams "back each other up" makes such suboptimalities much less disruptive than the "reverse prioritization" that can occur under pooling. Hence, altering the mix of patient types has a much more modest impact on performance in the streaming system. We relegate discussion of the model behind Figure 2.4 to Appendix B for the sake of brevity. We will test another important conjecture that streaming is more robust than pooling to misclassification errors in Section 2.6.

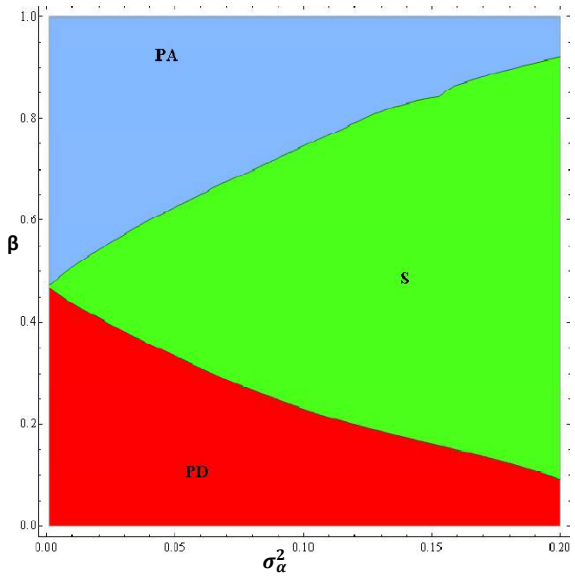


Figure 2.4: When the level of uncertainty in the percentage of A patients (measured by σ_α^2) increases, streaming becomes the optimal policy for an increasingly wide range of β values.

2.5 Phase 2 Implications of Streaming and Pooling

By modeling patient care as a single stage service process, the above model focused attention exclusively on Phase 1 sequencing. However, as we noted earlier, ED patients typically receive multiple visits from physician (designated as "treatment" states), interspersed with tests, waiting for test results and intermediate processing (designated as "wait" states), during which the patient is not available for interaction with the physician. To examine the Phase 2 sequencing decisions of which patient to see next whenever a physician completes a treatment stage, we now relax the single-stage service assumption and consider a multi-stage treatment process. Note that we still face the Phase 1 sequencing decision concerning the order in which to bring the patients back into the examination rooms. In both Phase 1 and Phase 2 sequencing, we can make use of ESI information, and, if available, A/D information. In Phase 2 sequencing, a physician can also potentially consider the number of past interactions with the patient. For instance, he/she could prioritize patients that have completed

more treatment stages because they may be closer to competition.

To explore the Phase 2 sequencing problem and its impact on the streaming vs. pooling comparison, we consider a static arrival (clearing) model with two physicians, where one physician is assigned to each stream under the streaming protocol. However, we represent the service process by the multi-stage model in Figure 2.5. In this model, after an initial wait state labeled as W1, patients go through an initial treatment (direct or indirect interaction with the physician) labeled as state T1 (so TTFT is the time between the start of T1 and the arrival of the patient). After T1, the patient oscillates between a stochastic number of "wait" (labeled as W) and "treatment" (labeled as T) states. We note that the treatment states start only if both the physician and the patient are available and the physician elects to work on that patient. After the final treatment by the physician, the patient experiences a final wait state (labeled as FW), which involves final processing by a nurse and a delay specific to admission (e.g., assignment to a bed) or discharge (e.g., final paper work and follow-up instructions), and then the patient exits the ED (to state E). To allow the distribution of physician interactions per patient to match observed data, we let the probability of a transition to the final stage (FW) depend on the number of previous interactions.

Because our focus here is on Phase 2 sequencing, we simplify some other aspects of the system to construct a tractable model. First, without loss of generality, we consider a single ESI level for patients. We do this, because in a clearing system, all ESI 2 patients will be served before ESI 3 patients (due to their Phase 1 sequencing priority). Hence, distinguishing between these patient classes will have little effect on system performance. Second, to permit maximum opportunity for Phase 2 sequencing, we assume there are enough examination rooms to hold all of the patients. Third, we assume that times in "wait" states (i.e., times spent for tests, waiting for test results and intermediate processing) are i.i.d. and exponentially distributed. For convenience, we also assume that times in the treatment states are i.i.d. and expo-

nentially distributed and are independent of the duration of wait states. The i.i.d. assumption glosses over any queueing for test equipment or nurses that could serve to correlate the times in the wait states. But since these states account for many different activities, we would not expect such correlation to be large. The exponential assumption reflects the unpredictability of the activities between physician interactions. Finally, to avoid the minor complications injected if preemption is disallowed, we allow preemption. For instance, when a patient returns from a test, the physician has the option of preempting the current patient and switching to the returning patient. We will relax these assumptions in the next section.

Because A and D patients have different performance metrics, it makes sense to treat them differently in Phase 2 sequencing. For D patients, LOS matters most. The work of [152] (which considers a manufacturing system with multiple phases of worker/product interaction) suggests that a "Pick-and-Run" policy can be effective when the performance criterion is average time spent in the system. Under this policy, the goal is to serve the most downstream job. In the ED, the equivalent policy would be for physicians to work on the patient closest to completion and try to complete this "old" patient's service (to the extent possible) before initiating a service for a "new" patient. We refer to this policy as *Prioritize Old (PO)*. In contrast, for A patients, TTFT is the key performance metric. Hence, for them, physicians should give preference to patients that have not yet been seen, unless constrained by the availability of exam rooms or the patient per physician limit. (Thus, in our simulation framework of the next section, where such additional constraints are also considered, a physician at his/her capacity should be directed to clear out in-process patients as quickly as possible by following the PO policy.) We refer to the policy that favors unseen patients as the *Prioritize New (PN)* policy.

We can show that these policies are optimal in the context of our simplified model (see Appendix A for a proof, where a Markov Decision Process is developed to analyze the underlying mutli armed restless bandit model).

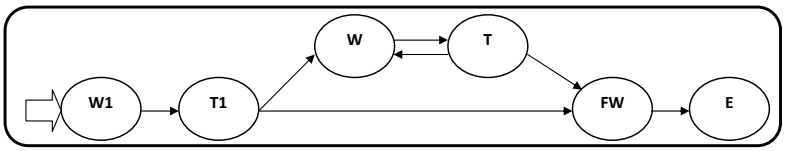


Figure 2.5: Multi-stage ED service: (W(1): (initial) wait, T(1):(initial) treatment, FW: final wait, E: exit).

Proposition 3 (Who to See Next?) *In the clearing model of a streaming ED flow with one physician assigned to each stream and multi-stage exponential treatment and wait stages modeled as in Figure 2.5:*
(i) If the probability of completion increases in the number of previous physician-patient interactions, the Prioritize Old (PO) rule is optimal (in the expected sense) for the D stream.
(ii) The Prioritize New (PN) rule is optimal (in the almost sure sense) for the A stream.

The implication of the above result is that instructing D physicians to work on the most downstream (old) patient and A physicians to work on the most upstream (new) patient should further improve the effectiveness of streaming. This addresses the *how* question we posed in Section 2.1 by suggesting a policy simple enough to be implemented in ED's. It also partially corresponds to what was done at Flinders (see [90]), where physicians assigned to the D stream were instructed that, in the absence of a threat to life/limb, need for time critical intervention or severe pain, they were to see patients in the order of arrival (i.e., a FCFS (First-Come-First-Served) mechanism). Moreover, "the staff were further encouraged to attempt as far as possible to complete one patient's journey before bringing the next patient out of the waiting room into a cubicle." However, physicians assigned to the A stream were instructed to continue to prioritize patients according to ESI categories and to use FCFS within each category. Our results suggest that Flinders sequencing policies within the ED are reasonable but not optimal.

We will confirm the conjecture that implementing the PO and PN rules for Phase 2 sequencing in the ED can enhance the effectiveness of streaming in the next section.

2.6 A Simulation-Based Comparison of Streaming and Pooling

We now test the conjectures suggested by our simple analytic models by means of a detailed simulation model of the ED. This simulation incorporates many realistic features discussed earlier, including dynamic non-stationary arrivals, multi-stage service, multiple physicians and exam rooms, inaccuracy in disposition prediction, bed-block by the hospital, among others. Our base case model was calibrated using a year of data from UMED plus time study data from the literature. Below, we highlight key features of the model. A more detailed description of our modeling assumptions is presented in Appendix C.

Patient Classes. As discussed earlier, patients are classified according to both ESI level (2 or 3) and ultimate disposition (A or D). This is done at the triage stage and results in patient classes 2A, 2D, 3A, and 3D. However, A/D prediction at triage is imperfect, resulting in misclassification errors. The true type of a patient is not revealed until the admit/discharge decision is made. Misclassification errors may vary from hospital to hospital but achievable levels seem to be in the range of $20 - 25\%$ ([90], [66], [94]).

Arrival Process. Arrivals for patient classes are modeled using non-stationary Poisson processes (which closely approximate the data) with arrival rates by class (obtained from a year of UMED data) depicted in Figure 3.5. The general pattern is similar to those reported in other studies (e.g., [54]).

Service Process. The ED service process is depicted in Figure 2.5. Each patient goes through several phases of patient-physician interactions/treatment followed by tests and preparations. The duration of each interaction is random and its average

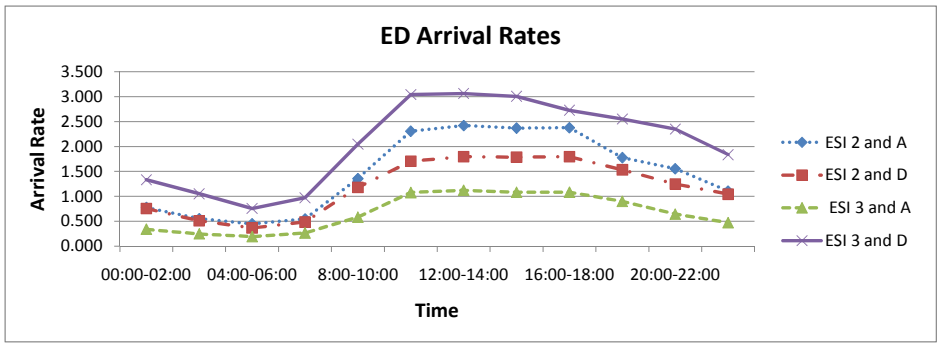


Figure 2.6: Class dependent arrival rates to the ED for an average day (obtained from a year of data in UMED).

may depend on the class of the patient and the number of previous interactions. For instance, the first and last interactions are usually longer than intermediate ones. The number of interactions with a physician ranges from 1 to 7 and depends on the class of the patient, as well as several other factors. Based on the class of the patient, we draw the number of such interactions from a distribution constructed from a detailed time study (see Table 3 of [52]) after modifying the data to represent our four patient classes (see Appendix C for details). The simulated service process is non-collaborative (an ED physician rarely transfers his/her patients to another physician) and non-preemptive (an ED physician rarely moves to another patient in the middle of his/her current interaction). The non-preemptive framework rules out impractical policies that for instance instruct physicians to visit each patient for a short time and then move to the other patient before finishing the interaction with the current patient. Such preemptive policies are generally avoided by physicians because they are inefficient for the physician (who will need to re-review patient history and condition upon the next return), as well as irritating to patients.

Physician-Patient Assignments. As noted earlier, the process of connecting patients with physicians involves two phases. In Phase 1, patients are assigned to available exam rooms, usually by the charge nurse, based on a Phase 1 sequencing rule. In Phase 2, whenever a physician becomes available, he/she chooses the next

patient (among those available/ready in the exam rooms) to see based on a Phase 2 sequencing rule. Under all patient flow designs, prioritizing ESI 2 patients over ESI 3 patients in Phase 1 is a constraint for safety reasons. For Phase 1 sequencing under streaming, patients are first streamed according to A/D information and then prioritized within streams with ESI 2 patients before ESI 3 patients (ties are broken with a FCFS rule). Under pooling, Phase 1 sequencing may or may not make use of A/D information, depending on the scenario under consideration. If A/D information is not available, then Phase 1 sequencing only considers ESI levels by prioritizing ESI 2 over ESI 3 with a FCFS rule to break ties. If A/D information is available under pooling, then Phase 1 sequencing prioritizes patients in the following order: 2A, 2D, 3A, 3D, with FCFS to break ties within a class.

In keeping with practice in UMED and elsewhere, we assume physicians do not take on more than seven patients at any time. We consider the following Phase 2 sequencing rules: (1) Service-In-Random-Order (SIRO), in which when a physician becomes available, s/he selects a patient at random from the pool of available (i.e., those not under a preparation or test) patients assigned to him and the new patients in the examination rooms waiting for a physician, provided that his/her total patient load does not exceed seven. This SIRO policy approximates current practice in many ED's in which physicians are not specifically encouraged to follow any specific rule, and hence, exogenous factors (changes in patient urgency level, patient discomfort, physician preference and experience, anticipation of interactions with testing facilities, access to newly available information, etc.) override systematic sequencing of patients. (We note, however, that while exogenous factors may make it appear that patients are sequenced according to SIRO, the decisions of physicians are not actually random. They are just based on criteria other than flow efficiency.) (2) First-Come-First-Served (FCFS), in which a physician selects his/her next patient in order of their arrival to the ED. This is an implementable policy to which many ED's aspire. (3) Prioritize-New-Prioritize-Old (PNPO), in which the Prioritize New (PN) policy is

used by physicians assigned to the A stream, and the Prioritize Old (PO) policy is used by physicians assigned to the D stream. That is, physicians in the A stream take an unassigned new patient whenever one is available in an exam room and the physician's patient load does not exceed seven. In contrast, physicians assigned to the D stream are instructed to prioritize the most down-stream patient assigned to them, in order to free up rooms and minimize LOS by completing patient journeys as quickly as possible. If a physician is handing seven patients s/he is asked to serve the most down-stream patient assigned to him regardless of the stream s/he is working in (in an effort to free up a room and lower his/her workload). Ties are always broken using a FCFS rule. While new to ED's, PNPO is also an implementable policy that our previous analytic results suggest should be effective.

Naming Convention. To distinguish between patient flow designs, we adopt a naming convention that labels each design as: Protocol/Phase 1/Phase 2. "Protocol" designates the type of system: pooling (P), streaming (S), and virtual streaming (VS). The difference between between the S and VS protocols is that S represents an implementation of streaming in which resources (rooms and physicians) are physically segregated and hence, idle resources assigned to one stream cannot be used by the patients of the other stream. In contrast, in VS, resources are only logically segregated and thus can be shared across streams. The "Phase 1" and "Phase 2" parts in the naming convention designate the Phase 1 and 2 sequencing rules described earlier. Phase 1 sequencing under streaming is done by separating patients based on their ultimate disposition (A or D) and prioritizing each stream by ESI level (2 before 3). Hence, we label all S and VS cases with "AD+ESI" to indicate the Phase 1 rule. Similarly, for "Phase 1" under pooling, we use "ESI" to denote the case where Phase 1 sequencing is based only on ESI information, and we use "AD+EDI" to denote the case where, in addition to ESI levels, A/D information is used to sequence patients in the order: 2A, 2D, 3A, 3D. For phase 2 sequencing rules, we use SIRO, FCFS, and PNPO. SIRO and FCFS can be used under either pooling or streaming, but PNPO

Table 2.1: Different patient flow designs under consideration and the notation implemented.

Protocol	Phase 1	Phase 2	Notation
Str. (S)	ESI only (ESI)	Service In Random Order (SIRO)	S/ESI/SIRO
		First Come First Served (FCFS)	S/ESI/FCFS
		Prioritize New Prioritize Old (PNPO)	S/ESI/PNPO
	A/D Info + ESI (AD+ESI)	Service In Random Order (SIRO)	S/AD+ESI/SIRO
		First Come First Served (FCFS)	S/AD+ESI/FCFS
		Prioritize New Prioritize Old (PNPO)	S/AD+ESI/PNPO
Pooling (P)	ESI only (ESI)	Service In Random Order (SIRO)	P/ESI/SIRO
		First Come First Served (FCFS)	P/ESI/FCFS
	A/D Info + ESI (AD+ESI)	Service In Random Order (SIRO)	P/AD+ESI/SIRO
		First Come First Served (FCFS)	P/AD+ESI/FCFS
Vir. Str. (VS)	ESI only (ESI)	Service In Random Order (SIRO)	VS/ESI/SIRO
		First Come First Served (FCFS)	VS/ESI/FCFS
		Prioritize New Prioritize Old (PNPO)	VS/ESI/PNPO
	A/D Info + ESI (AD+ESI)	Service In Random Order (SIRO)	VS/AD+ESI/SIRO
		First Come First Served (FCFS)	VS/AD+ESI/FCFS
		Prioritize New Prioritize Old (PNPO)	VS/AD+ESI/PNPO

can only be implemented in S and VS systems where physicians and patient classes are segregated into A and D streams. Table 2.1 summarizes this notation and the possible patient flow designs.

By comparing the pooling and streaming policies (S and VS) under the basic SIRO Phase 2 sequencing rules, we can address the question of *whether* streaming can improve ED performance. By performing sensitivity analysis on the model parameters, we address the question of *where* streaming is most effective. And by matching streaming and pooling with various Phase 1 and Phase 2 sequencing rules, we gain insight into the question of *how* to implement streaming for maximum effectiveness.

In the following subsections, we present our main findings from the simulation experiments. For each patient flow design described above, the objective function $(\beta TTFT(A) + (1-\beta)LOS(D))$ is computed as an average over 5000 replications of a week of operation, where the result for each replication is obtained after a warmup period of one week. Further details about our simulation framework can be found in Appendix C.

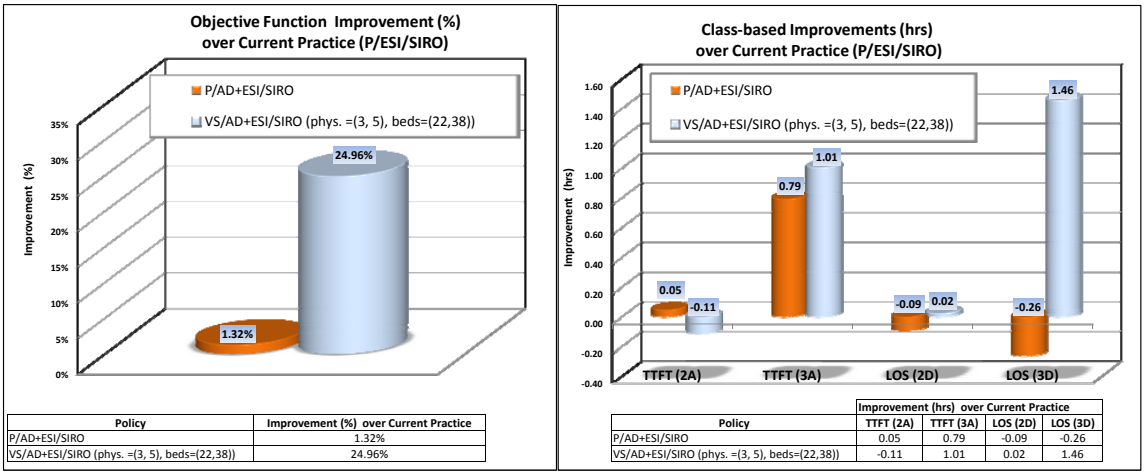
2.6.1 ED Flow Design: Pooling, Physical Streaming, or Virtual Streaming?

We start with a comparison between the current practice of pooling in the ED's and physical streaming (where, unlike virtual streaming and our analytical clearing model, capacity sharing is not possible).

Observation 2 *Comparing simulations of the S/AD+ESI/SIRO and P/ESI/SIRO systems shows that pooling is more effective than physical streaming, with a 77% lower objective value.*

The inefficiency of physical streaming results from the imbalanced and low utilization of resources (which leads to intervals in which physicians are starved for lack of a patient or bed, even though a patient and/or bed is available in the opposite stream). In other words, physical streaming exhibits an "anti-pooling effect," which occurs because physical separation in either physicians or beds prevents capacity sharing. To place the observed magnitude of the anti-pooling effect of physical streaming (77%) in context, we make use of Kingman's formula for a $G/G/s$ queueing system with $s = 8$ physicians and two parallel $G/G/s$ queueing systems with $s = 4$ physicians each, with a server utilization close to our base case. The pooling benefit of having a $G/G/8$ queue versus two parallel $G/G/4$ queues on the average waiting time and the average system time is 79% and 7%, respectively. Since our objective function is a weighted average of TTFT (queue time) and LOS (system time), we would expect the anti-pooling penalty to fall between these values, as it does. This simple example illustrates that, even when capacity is perfectly balanced, the inability to share capacity between streams can be very damaging to performance. In the ED, this effect is particularly pronounced (i.e., toward the higher end of the range indicated by the G/G/s model) because (1) it is not possible to balance utilization in the two streams exactly due to the discreteness of physicians and beds, and the fact that the average mix of A and D patients fluctuates according to the time of day (see Figure 7), and (2)

the limited number of beds in the ED means that patients can be held in the waiting room even when physicians are idle, an effect that becomes more pronounced when beds are separated into two smaller systems under physical streaming. (The magnitude of this effect becomes apparent when we observe that the anti-pooling penalty falls to 17% in the simulation model when the number of beds is made infinite.) As a result, physical streaming is decidedly worse for performance than is a conventional pooling protocol. This leads us to suspect that Flinders does not rigidly adhere to a complete physical separation of streams, even though they described their system as such.



Policy	Improvement (%) over Current Practice
P/AD+ESI/SIRO	1.32%
VS/AD+ESI/SIRO (phys. =(3, 5), beds=(22,38))	24.96%

	Improvement (hrs) over Current Practice			
Policy	TTFT (2A)	TTFT (3A)	LOS (2D)	LOS (3D)
P/AD+ESI/SIRO	0.05	0.79	-0.09	-0.26
VS/AD+ESI/SIRO (phys. =(3, 5), beds=(22,38))	-0.11	1.01	0.02	1.46

Figure 2.7: Virtual streaming significantly outperforms pooling and improved pooling. The reason is that VS dramatically decreases LOS for 3D patients with only a minor increase in TTFT of 2A patients. (Results for an ED with 8 physicians and 60 beds, a 20% misclassification error rate, and a weight for TTFT of A patients of $\beta = 0.50$).

Since physical streaming is so unattractive, we do not consider it further and instead we investigate whether virtual streaming (VS) can improve ED performance. We start by considering the SIRO Phase 2 sequencing rule (as an approximation of the status quo in most ED's) and compare VS/AD+ESI/SIRO (basic virtual streaming) and P/AD+ESI/SIRO (improved pooling) with P/ESI/SIRO (current pooling practice in most ED's). Figure 2.7 depicts the simulation results. The graph on the left depicts the percentage improvement in the combined objective function (with

$\beta = 0.5$). The graph on the right illustrates the improvement (in hours) achieved for each class of patients separately. The significant improvement shown in Figure 2.7 (left) is achieved because VS dramatically decreases LOS of 3D's while only slightly increasing TTFT of 2A's (see Figure 2.7 (right)).

Observation 3 *Virtual streaming significantly outperforms both pooling and improved pooling by striking a better balance between TTFT of A's and LOS of D's.*

Since VS does not require any physical reconfiguration of the ED, this finding provides strong evidence that virtual streaming can be an attractive and practical option for improving ED responsiveness. As can be easily observed from Figure 2.7 (right), this attractiveness is also very robust to the weights assigned to our two main metrics, TTFT for A's and LOS for D's.

To further confirm this insight, we also compare the performance of the proposed virtual streaming (VS/AD+ESI/PNPO) with the current practice (P/ESI/SIRO) using all four metrics (i.e., TTFT and LOS for both A's and D's). Table 2.2 presents these four metrics in hours for our base case scenario under pooling and streaming. To examine the robustness of streaming, we consider a weighted average of all these four metrics defined as $TTFT(A)+\beta_1 TTFT(D)+\beta_2 LOS(A)+\beta_3 LOS(D)$, where the weight for $TTFT(A)$ is assumed to be one and other weights represent the relative priorities of the remaining metrics to that of $TTFT(A)$. Our analysis reveals that pooling is only preferred in unrealistic cases where (a) almost no weight is placed on $LOS(D)$ (i.e., β_3 is small), (b) $LOS(A)$ is weighted more heavily than $TTFT(A)$ (i.e., $\beta_2 > 1$), and (c) $LOS(A)$ is more heavily weighted than $TTFT(D)$ (i.e., $\beta_2 > \beta_1$). Condition (a) is problematic, since (as we discussed previously) $LOS(D)$ is of great concern for ED's. Conditions (b) and (c) are particularly unrealistic because A's remain in the hospital well beyond their stay in the ED, and hence, LOS in the ED is not that important for them. These provide further evidence that (1) the benefit of the proposed streaming policy (over the current pooling policy) is robust with respect

to weights assigned, and (2) considering an objective function made up of the two most important metrics, $TTFT(A)$ and $LOS(D)$, is a reasonable approximation of the full multi-objective optimization problem. Hence, for the reminder of of our analyses, we will make use of the two dimensional objective function involving only $TTFT(A)$ and $LOS(D)$. However, it is worth noting that, based on the results presented in Table 2.2, we also expect the percentage of left without being seen (LWBS) metric to be improved by the proposed streaming design, as it improves the TTFT of both A's and D's.

Since patients who abandon the ED are not tracked in detail, we do not have enough data (e.g., how long they waited before leaving) to characterize the exact effect of streaming on LWBS. However, we can get an estimate using the following method. First, we assume patients may leave after an exponentially distributed amount of time if they are not yet seen. This is a reasonable approximation of reality if there are multiple factors leading to a patient abandonment, each occurring according to a Poisson process. Under these conditions, the patient abandonment process is a superposition of Poisson processes which is itself Poisson. To estimate the rate of this process, we note that the current LWBS percentage in the UMED is 3%. Moreover, based on Table 2.2, the TTFT for an average patient (A or D) is about 1 hour. Thus, we need to find the exponential distribution that has a cdf equal to 0.03 at $TTFT = 1$. This leads to an exponential distribution with rate 0.031. Next, augmenting the arrival rates in the simulation by the current percentage of LWBS, 3%, and having patients leave after this exponential time, we observe that the LWBS (when made endogenous) under the streaming scenario is 1.04% compared to that of 3% in the current pooling system. Because the LWBS is reduced, the arrival rate to the ED is increased which in turn slightly increases the TTFT relative to what it would be without the LWBS improvement. Nevertheless, streaming still significantly improves TTFT compared to current pooling practice in addition to achieving a significant reduction in the percentage LWBS. The bottom line is that streaming can reduce

Table 2.2: Performance (in hrs) of the proposed streaming design (VS/AD+ESI/PNPO) and current pooling practice (P/ESI/SIRO) under four metrics as well as the associated LWBS (%). For the streaming design the physician and bed split have been optimized at phys. =(3, 5) and beds=(22,38) for the A and D sides, respectively.

Policy	**TTFT (A)**	**TTFT (D)**	**LOS (A)**	**LOS (D)**	**LWBS**
P/ESI/SIRO	0.88532	1.07893	7.4458	3.51401	3%
VS/AD+ESI/PNPO (exogenous LWBS)	0.67253	0.95437	7.7389	2.60942	3%
VS/AD+ESI/PNPO (endogenous LWBS)	0.74601	1.01349	7.8134	2.67707	1.4%

overall TTFT, LOS and LWBS relative to pooling. However, as illustrated in Table 2.2, it does this by allowing a slight increase in LOS for A patients in order to achieve substantial improvements in all other metrics.

Having answered the *whether* question, we now seek to answer *how* VS should be implemented for maximum benefit. Proposition 3 suggests that following the PNPO rule for Phase 2 sequencing may further improve performance. Using our simulation test bed we observe that this conjecture is true. However, we also observe that improved Phase 2 sequencing does not make as large an improvement as that achieved by virtual streaming.

Observation 4 *Using the PNPO rule for Phase 2 sequencing improves the performance, but performance of VS is relatively insensitive to the Phase 2 sequencing rule, indicating that most of the benefit of streaming is attributable to Phase 1.*

The insensitivity of performance to Phase 2 sequencing is due to the fact that ED physicians frequently do not have many patients to choose among, because patients are often unavailable while waiting for test results. In ED's with shorter test times, higher physician utilization, and larger case loads (patients per physician), there would be more choice among in-process patients, and hence more benefit from an improved Phase 2 sequencing policy.

To get a sense of the maximum achievable value of the PNPO policy, we considered an ED with 50% shorter test times than UMED, as well as higher maximum case loads

(12 vs. 7) and very high dedicated utilization (up to 88% compared to 44% in the base case). "Dedicated utilization" refers to the fraction of the time that a physician is involved in activities that will not be interrupted to see another patient. These include direct care of patients and some indirect activities (e.g., reading patient test results). But physicians also engage in many indirect activities (e.g., staff management, paper work, discussions with colleagues) that are preemptible and hence do not contribute to patient queueing. Studies report that direct care activities occupy 32% of ED physician time ([68]), so the 44% value for dedicated utilization in our base model is plausible. Of course, total ED physician utilization, which includes all direct and indirect activities is much higher; ED physicians are busy. But here we are only concerned with dedicated utilization, since this is what drives congestion.

The percentage improvement due to implementing the PNPO policy is shown in Figure 3.4 for a range of dedicated utilization values. This figure confirms that implementing PNPO becomes more effective when (1) the dedicated utilization of physicians is high, (2) the number of patients allowed per physician is large, and (3) patient test times are short. This suggests a practical limit of 4% on the amount of improvement possible via better Phase 2 sequencing. When combined with the benefit of virtual streaming, this results in a 29% improvement in the overall objective function compared to the current pooling practice (P/ESI/SIRO).

2.6.2 Sensitivity Analyses: Where to Implement Virtual Streaming?

Having addressed the *whether* and *how* questions we raised in Section 2.1, we now turn to the question of *where* virtual streaming is likely to be most attractive. We address this by performing sensitivity analyses on environmental characteristics in order to identify key factors that amplify the advantage of implementing virtual streaming over pooling.

To this end, in addition to using V/AD+ESI/PNPO as a good candidate for vir-

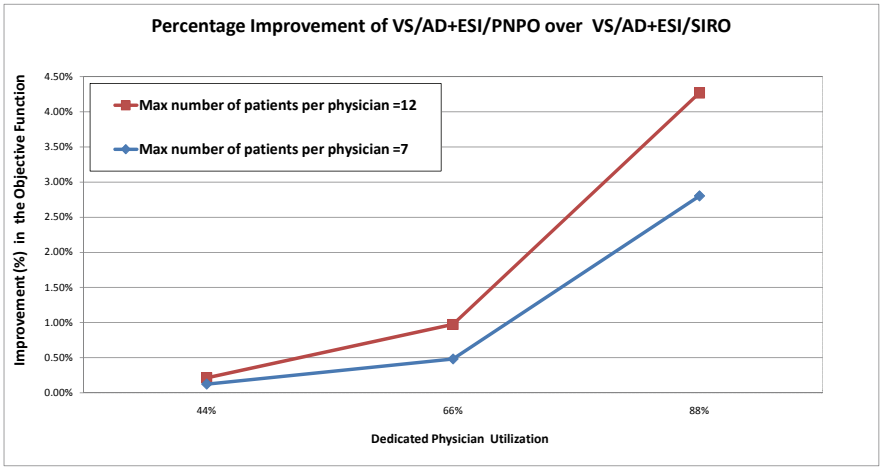


Figure 2.8: The benefit of implementing PNPO sequencing rule. ED's with a higher physician utilization or with a higher maximum number of patients allowed per physician benefit more from PNPO.

tual streaming, we select P/AD+ESI/FCFS as a good candidate for pooling because: (a) it makes use of A/D information in Phase 1 sequencing, and (b) FCFS is an implementable policy, which was used at Flinders, and showed a small improvement over SIRO for Phase 2 sequencing in our simulation experiments. However, as we observed previously, the effect of a Phase 2 sequencing rule is small compared to the benefit obtained from virtual streaming, so we do not expect the results to be sensitive to the Phase 2 sequencing rule.

We start by examining the role of misclassification errors and β (the relative weight given to TTFT of A patients compared to LOS of D patients) on the relative benefit of virtual streaming over pooling. Based on our earlier clearing model, we conjectured that a higher β should favor pooling. Common sense suggests that A/D information is less valuable if it is inaccurate, so we also expect a higher misclassification probability to also favor pooling. Figure 2.9 confirms these conjectures and shows that unless an ED gives an extremely very heavy weight to the TTFT of A patients (high β) or has a very high misclassification error rate, virtual streaming is preferred to pooling.

Next we consider the effect of the percentage of A patients (α). Our analytical model in Section 2.4 led us to conjecture that a higher mean or a higher day-to-

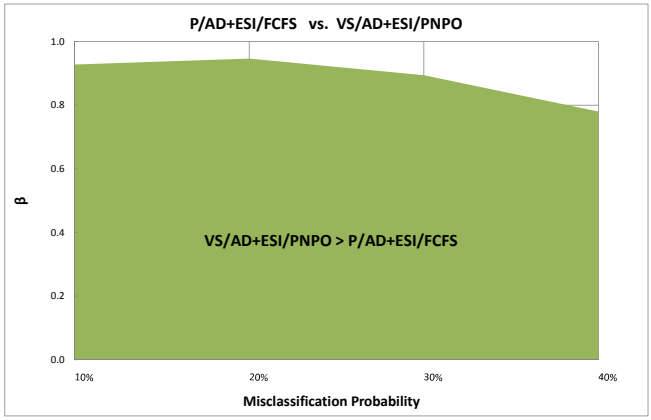


Figure 2.9: Sensitivity of virtual streaming and pooling designs. Lower weight on TTFT of A patients (β) or misclassification probability favors virtual streaming over pooling.

day variance in the percentage of A patients increases the attractiveness of virtual streaming. Figure 2.10 (left) shows simulation results indicating that virtual streaming is indeed more attractive in ED's with a higher percentage of A patients. Figure 2.10 (right) shows the effect of increasing day-to-day variation in the mix of patients by drawing α from a family of beta distributions, $\text{Beta}\big(f(x), 2f(x)\big)$ where $f(x) = (2 - 9x)/(27x)$, $x \in (0, 2/9)$. (Recall that doing this holds the mean at $\mu_\alpha = 1/3$ (which approximates UMED data), but allows the variance, $\sigma_\alpha^2 = x$, to range from 0 to 2/9.) This indicates that higher variability in α also makes virtual streaming more attractive, as our analytic models predicted.

Observation 5 *A higher fraction of A patients and a higher variance in the day-to-day fraction of A patients both favor (virtual) streaming relative to pooling.*

It is worth nothing that the percentage of A patients at Flinders is relatively high ($\alpha = 43\%$) compared to the average rate of admission in the U.S. ED's, which was $\alpha = 12.8\%$ in 2006 (NHAMCS by [114]). This may be one reason that streaming was considered a success at Flinders.

Another environmental factor that affects the (virtual) streaming versus pooling comparison is the relative test and treatment times of A's versus D's. In Figure 2.11,

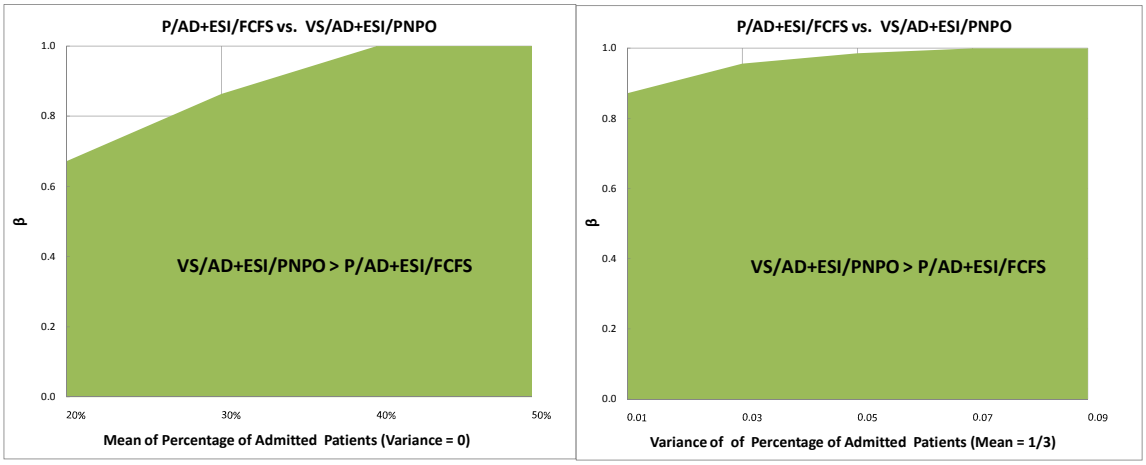


Figure 2.10: Sensitivity of virtual streaming (VS) and pooling (P) designs with respect to mean μ_α (left) and variance σ_α^2 (right) of the percentage of A patients. VS dominates P in the shaded region.

we examine the sensitivity of the VS/AD+ESI/PNPO and P/AD+ESI/FCFS configurations to increases in the test times of A (left) and D (right) patients. In Figure 2.12, we similarly consider the sensitivity of these two configurations to increases in the treatment times of A (left) and D (right) patients.

Observation 6 *Increasing the difference between the test and/or treatment times of A and D patients increases the attractiveness of virtual streaming relative to pooling.*

This observation has potentially important consequences for where virtual streaming is likely to be effective. First, ED's with congested or slow test facilities (which are used more frequently by A's than by D's) are likely to benefit more from virtual streaming than ED's with fast or ample test facilities. Second, ED's that handle serious/complex patients among their A's (e.g., Level 1 trauma centers and teaching hospitals) are more likely to benefit from virtual streaming than ED's with less extreme A's (e.g., community hospitals), because the former is likely to have a larger gap between treatment times of A's and D's.

To further answer the *where* question, we consider the impact of a common phenomenon in ED's, the so-called "bed-block" process, which occurs when A patients

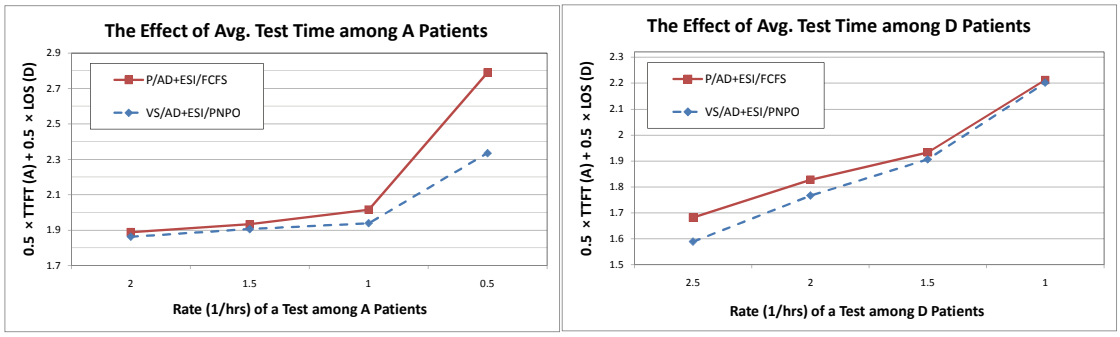


Figure 2.11: The effect of average patient test time (MRI, CT Scan, etc.) on the relative performance of two virtual streaming and pooling configurations. As test time for A patients increases (left) or decreases for D patients (right), virtual streaming becomes more attractive compared to pooling.

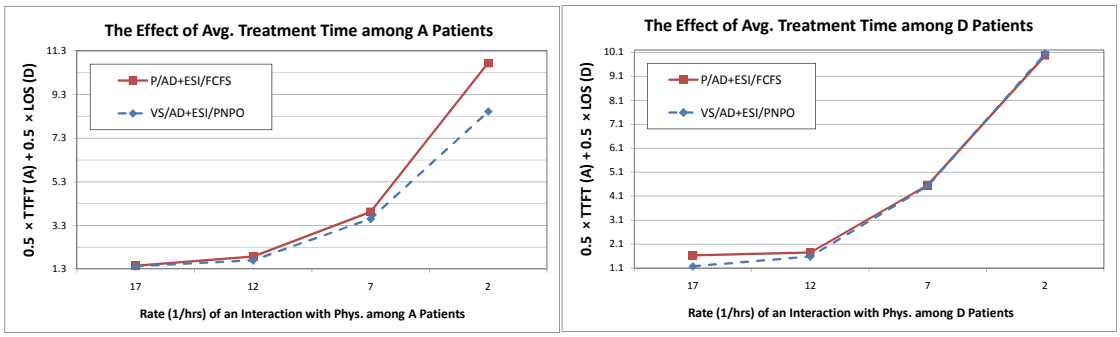


Figure 2.12: The effect of treatment times on the relative performance of two virtual streaming and pooling configurations. As treatment time for A patients increases (left) or decreases for D patients (right), virtual streaming becomes more attractive compared to pooling.

are boarded in the ED while they wait for a hospital bed. Decreasing bed-block times has been shown to be one of the most significant factors (even more significant than increasing the number of beds) in reducing LOS ([88]). However, its impact on streaming has not been studied. Figure 2.13 compares the performance of the VS/AD+ESI/PNPO and P/AD+ESI/FCFS configurations for various values of the average boarding time of an A patient.

Observation 7 *The relative attractiveness of virtual streaming over pooling increases with the average boarding time of A patients.*

The implication is that ED's with higher frequency of bed-block or longer waits for hospital beds can benefit more from virtual streaming.

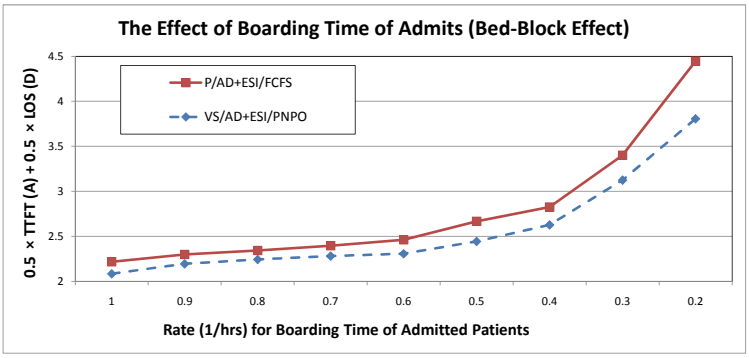


Figure 2.13: The effect of the average boarding time on the performance of two virtual streaming and pooling configurations. ED's with longer boarding of A's benefit more from virtual streaming.

Finally, we consider the effect of the average dedicated utilization of physicians on the attractiveness of virtual streaming. Figure 2.14 (left) depicts the objective function for policies VS/AD+ESI/PNPO and P/AD+ESI/FCFS, while Figure 2.14 (right) shows the improvement in the objective function from implementing VS/AD+ESI/PNPO instead of P/AD+ESI/FCFS.

Observation 8 *The relative attractiveness of virtual streaming over pooling increases with average dedicated utilization of physicians.*

The implication is that congested ED's with high arrival rates or a low number of physicians can benefit more from virtual streaming. Furthermore, we did not explicitly account for physician interruptions, such as treating ESI-1 patients or dealing with other non-patient issues, which would add to physician's non-preemptible activities (and hence dedicated utilization). Thus, our estimates of the benefits of virtual streaming are probably conservative.

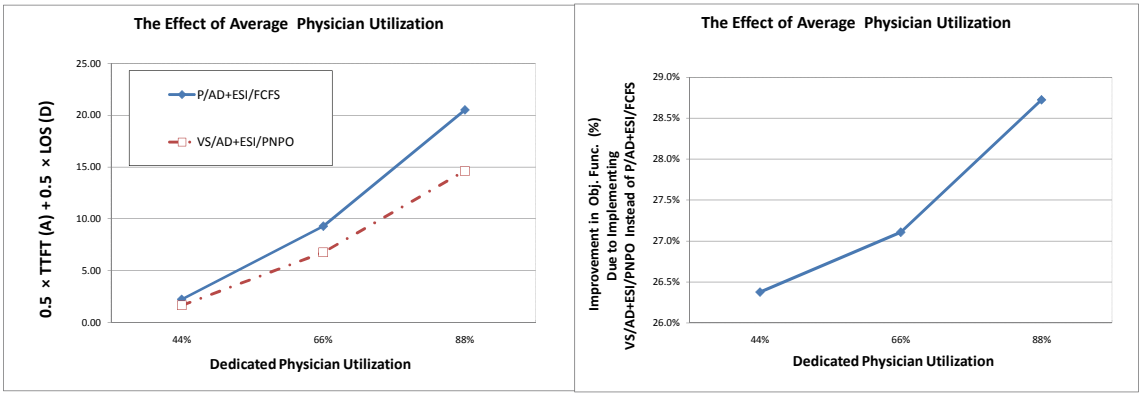


Figure 2.14: The effect of average physician utilization on the attractiveness of virtual streaming. ED's with higher average physician utilization benefit more.

2.7 Conclusion

This chapter describes our investigation of a new approach to managing patient flows in ED's: streaming, which separates patients based on an up-front prediction on their final disposition (admission or discharge). Streaming has been popularized by Flinders Medical Center, where it has been credited with dramatically reducing patient length of stay (LOS). While the empirical results reported by Flinders have stimulated substantial interest among ED professional, they are not conclusive because (1) the Flinders experiment was not a controlled study, so a Hawthorne effect cannot be ruled out, (2) other changes (e.g., lean) were implemented along with streaming, and (3) the environment at Flinders may not reflect other ED's (e.g., the fraction of A patients at Flinders is substantially above the norm). Indeed, our results suggest that the physical streaming approach as described by the Flinders may actually degrade ED performance because of an "anti-pooling" effect caused by separating resources into segments. Hence, we suspect that the Flinders success is partly due to informal capacity sharing to overcome the anti-pooling effect and partly due to other process improvements.

To avoid the anti-pooling effect of physical streaming, we proposed virtual streaming, in which physicians and rooms are only logically separated and, hence, excess

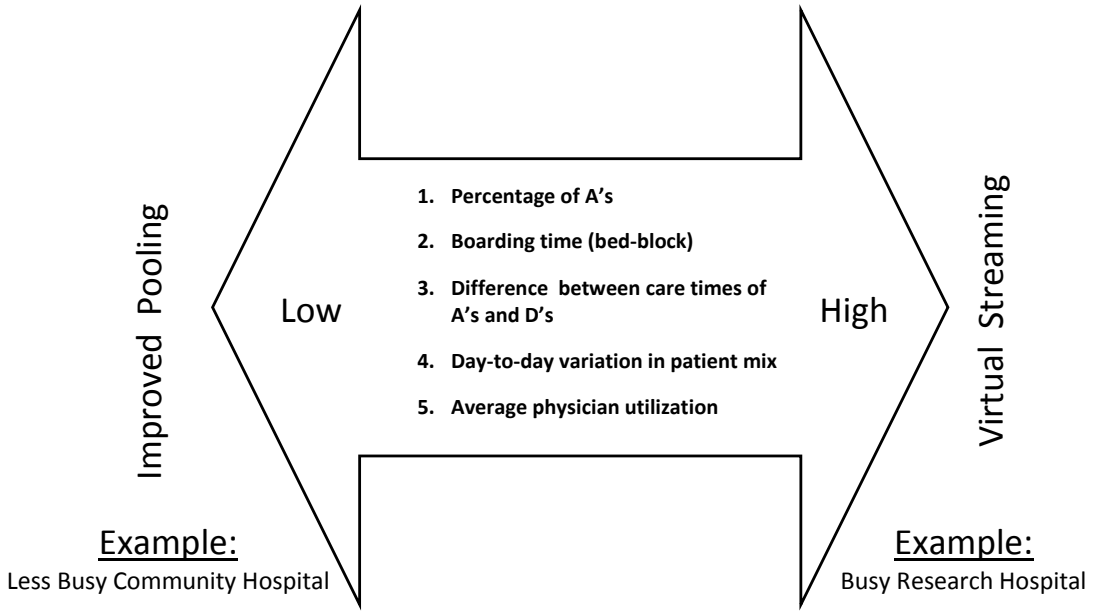


Figure 2.15: ED patient flow design strategy based on key environmental characteristics of the ED.

capacities can be shared. Using simple analytical models, we found that virtual streaming can strike a better balance between the TTFT of A patients and the LOS of D patients by devoting some capacity to each patient type, rather than giving full priority to one. These analytic models also led to several conjectures about the environmental factors that should make virtual streaming more attractive.

We tested these conjectures with a realistic simulation and found that virtual streaming can indeed significantly improve ED performance (by 25% in a case designed to represent the ED of a busy academic hospital). Since implementing virtual streaming does not require a physical layout redesign in the ED, it provides a practical option to improve ED responsiveness.

We also found that the information used to stream patients (i.e., A or D classification) can be used by physicians to sequence patients within exam rooms and achieve additional performance improvements (up to 4% beyond the improvement due to virtual streaming alone). To achieve this, physicians assigned to the A stream should use (to the extent possible) a "Prioritize New" rule that favors seeing new

patients before finishing patients already in progress, while physicians assigned to the D stream should use (to the extent possible) a "Prioritize Old" rule that favors completing patient journeys before initializing new ones.

Our results also indicate that while virtual streaming can be effective, it is not uniformly attractive to all ED's. Figure 2.15 summarizes the results of our sensitivity analyses, which suggest that virtual streaming is best suited for ED's with (1) a high percentage of A patients, (2) longer service times for A's than D's, (3) long patient boarding times due to bed-block, (4) high day-to-day variations in patient mix, and (5) high average physician utilization. Using a PNPO Phase 2 sequencing rule is more effective in ED's with (1) high average physician utilization, (2) large patient case load, and (3) short waits for test results.

In broad terms, our results indicate that better triage information about patients (e.g., A/D classification) can be leveraged to improve ED performance. One question to be answered in future research is whether other types of pre-treatment information (e.g., case complexity, type of testing required, etc.) are possible to obtain and yield additional benefit. Given the crisis levels of ED congestion, it is critical to find out.

2.8 Appendix A: Proofs.

Proof of Proposition 1. We use a sample path argument. Consider the probability space $(\Omega, \mathcal{F}, \mathcal{P})$. Let $CA_k^\pi(\omega)$ and $CD_k^\pi(\omega)$ denote the completion time of kth Admit and kth Discharge type patient (under policy π and along sample path $\omega \in \Omega$), respectively. Also, assume $T_A^\pi(\alpha, \omega)$ and $L_D(\alpha, \omega)$ denote the (average) TTFT of Admits and the (average) LOS of Discharges for a given $\alpha \in [0, 1]$ and sample path $\omega \in \Omega$, respectively.

Proof of Part (i). To prove part (i), it is sufficient to show that for every α and every sample path ω: (a) $T_A^{PA}(\alpha, \omega) \leq T_A^S(\alpha, \omega)$, and (b) $T_A^{PA}(\alpha, \omega) \leq T_A^{PD}(\alpha, \omega)$. To prove (a), fix α and let $t(\omega) = \min\{CA_{n_A}^S(\omega), CD_{n_D}^S(\omega)\}$ denote the time that

system moves to a pooling scenario under Streaming policy and over sample path ω. If $t(\omega) = CA^S_{n_A}(\omega)$ (i.e., if Streaming becomes Pooling when Admits are all served) then notice that under $\pi = S$, the kth Admit patient starts its treatment at $CA^S_{k-1}(\omega)$ but under $\pi = PA$, the kth Admit patient starts its treatment at $\min\{CA^{PA}_{k-1}(\omega), CA^{PA}_{k-2}(\omega)\} \leq \min\{CA^S_{k-1}(\omega), CA^S_{k-2}(\omega)\} \leq CA^S_{k-1}(\omega)$, where the first inequality can be easily shown using induction on k, and the second inequality trivially holds. Hence, under $\pi = PA$ each patient is seen no later than when s/he is seen under $\pi = S$, and therefore (a) holds. Now if $t(\omega) = CD^S_{n_D}(\omega)$ (i.e., if Streaming becomes Pooling when some Admits still have not been seen), assume the last Admit type patient that has been seen before or at time $t(\omega)$ under $\pi = S$ is the $n_t(\omega)$th patients of this type. Using the previous argument, none of first $n_t(\omega)$ patients under $\pi = S$ are seen before the time they would have been seen under $\pi = PA$. Moreover, under $\pi = S$ every remaining Admit patient is seen with a constant delay of at least $t(\omega) - CA^{PA}_{n_t(\omega)-1}(\omega) \geq 0$ compared to what it would have been seen under $\pi = PA$. Therefore, for every ω and every α, every Admit type patient is seen under $\pi = S$ no sooner than what it would have been seen under $\pi = PA$. Thus (a) holds. To show (b), fix α and notice that under $\pi = PD$ every Admit patient is seen with a constant delay of at least $CD^{PD}_{n_D-1}(\omega)$ compared to what it would have been seen under $\pi = PA$. Thus, (b) holds and the proof of (i) is complete.

Proof of Part (ii). To prove part (ii), it is sufficient to show that for every α and every sample path ω: (1) $L^{PD}_D(\alpha, \omega) \leq L^S_D(\alpha, \omega)$, and (2) $L^{PD}_D(\alpha, \omega) \leq L^{PA}_D(\alpha, \omega)$. To show (1), fixing α, we show that $CD^{PD}_k(\omega) \leq CD^S_k(\omega)$ $(\forall k \in 1, 2, \cdots, n_D)$. To show this notice that using the same argument as part (i) (and after swapping labels D and A) it is easy to show that TTFT of each Discharge patient under $\pi = PD$ is no more than its TTFT under $\pi = S$. That is, if $TD^\pi_k(\omega)$ denotes the TTFT of the kth Discharge patient under sample path ω, then $TD^{PD}_k(\omega) \leq TD^S_k(\omega)$ $(\forall k \in 1, 2, \cdots, n_D)$. Next, if $SD_k(\omega)$ is the service time of kth Discharge patient under sample path ω, $CD^\pi_k(\omega) = TD^\pi_k(\omega) + SD_k(\omega)$. Thus, since $TD^{PD}_k(\omega) \leq TD^S_k(\omega)$, we

have $CD_k^{PD}(\omega) \leq CD_k^S(\omega)$ $(\forall k \in 1, 2, \cdots, n_D)$, and hence (1) holds. To show (2), fix α and notice that the completion time of every Discharge patient under PA is delayed at least for $CD_{n_A-1}^{PA}$ units of time compared to PD, and hence, the proof is complete. □

Proof of Lemma 1. To prove this lemma, using the definition of β-convexity, we need to show that sets $\mathcal{A}^\pi$ $(\forall \pi \in \mathbf{\Pi})$ are convex in β for every α. Fix α and consider β_1 and β_2 such that $(\alpha, \beta_1) \in \mathcal{A}^\pi$ and $(\alpha, \beta_2) \in \mathcal{A}^\pi$. We then need to show that $(\alpha, \gamma\beta_1 + (1-\gamma)\beta_2) \in \mathcal{A}^\pi$ for every $\gamma \in [0, 1]$. Notice that as $(\alpha, \beta_1) \in \mathcal{A}^\pi$, for every other policy $\pi' \in \mathbf{\Pi}$ we have:

$$\beta_1 T_A^\pi(\alpha) + (1-\beta_1) L_D^\pi(\alpha) \leq \beta_1 T_A^{\pi'}(\alpha) + (1-\beta_1) L_D^{\pi'}(\alpha). \tag{2.1}$$

Similarly, as $(\alpha, \beta_2) \in \mathcal{A}^\pi$, for every other policy $\pi' \in \mathbf{\Pi}$ we have:

$$\beta_2 T_A^\pi(\alpha) + (1-\beta_2) L_D^\pi(\alpha) \leq \beta_2 T_A^{\pi'}(\alpha) + (1-\beta_2) L_D^{\pi'}(\alpha). \tag{2.2}$$

Now multiplying both sides of (2.1) by γ and both sides of (2.2) by $(1-\gamma)$ and adding up the resulting inequalities we get:

$$\begin{aligned}
&(\gamma\beta_1 + (1-\gamma)\beta_2)\, T_A^\pi(\alpha) + (1 - [\gamma\beta_1 + (1-\gamma)\beta_2]) L_D^\pi(\alpha) \\
&\leq (\gamma\beta_1 + (1-\gamma)\beta_2)\, T_A^{\pi'}(\alpha) + (1 - [\gamma\beta_1 + (1-\gamma)\beta_2]) L_D^{\pi'}(\alpha).
\end{aligned}$$

Hence, since the above inequality holds for every $\pi' \in \mathbf{\Pi}$ and every $\gamma \in [0, 1]$, $(\alpha, \gamma\beta_1 + (1-\gamma)\beta_2) \in \mathcal{A}^\pi$ for every $\gamma \in [0, 1]$. Thus, the optimal strategy $\pi^*(\alpha, \beta)$ is convex in β. □

Proof of Proposition 2. Define functions $\beta_1(\alpha)$ and $\beta_2(\alpha)$ as follows:

$$\begin{aligned}
\beta_1(\alpha) &= \inf\{\beta : f^S(\alpha, \beta) \leq f^{PD}(\alpha, \beta)\}, \\
\beta_2(\alpha) &= \sup\{\beta : f^S(\alpha, \beta) \leq f^{PA}(\alpha, \beta)\}.
\end{aligned}$$

We show that by setting $\underline{\beta}(\alpha) = \min\{\beta_1(\alpha), \beta_2(\alpha)\}$ and $\bar{\beta}(\alpha) = \max\{\beta_1(\alpha), \beta_2(\alpha)\}$, Streaming is optimal for a given α if, and only if, $\beta(\alpha) \in [\underline{\beta}(\alpha), \bar{\beta}(\alpha)]$. To see the "if" part, fix α, suppose $\beta(\alpha) \in [\underline{\beta}(\alpha), \bar{\beta}(\alpha)]$, and write $\beta(\alpha)$ as a convex combination of extreme points $\underline{\beta}(\alpha)$ and $\bar{\beta}(\alpha)$. Then notice that by definition of $\underline{\beta}(\alpha)$ and $\bar{\beta}(\alpha)$, Streaming is optimal at both extreme points $\underline{\beta}(\alpha)$ and $\bar{\beta}(\alpha)$. Hence, by Lemma 1 Streaming is also optimal at $\beta(\alpha)$. To see the "only if" part, fix α and suppose $\beta(\alpha) \notin [\underline{\beta}(\alpha), \bar{\beta}(\alpha)]$. That is, suppose for some $\epsilon > 0$ either (a) $0 \leq \beta(\alpha) \leq \underline{\beta}(\alpha) - \epsilon$, or (b) $\bar{\beta}(\alpha) + \epsilon \leq \beta(\alpha) \leq 1$. If (a) holds, write $\beta(\alpha)$ as a convex combination of $\tilde{\beta}(\alpha) = 0$ and $\underline{\beta}(\alpha) - \epsilon$. Then notice that, from Proposition 1, $\pi = PD$ is optimal at $\tilde{\beta}(\alpha) = 0$. Also, $\underline{\beta}(\alpha) - \epsilon < \underline{\beta}(\alpha) \leq \beta_1(\alpha)$. Therefore, from the definition of $\beta_1(\alpha)$, $\pi = PD$ is better than $\pi = S$ at $\underline{\beta}(\alpha) - \epsilon$. Moreover, $\pi = PA$ cannot be optimal at $\underline{\beta}(\alpha) - \epsilon$, since otherwise, choosing a β in $[\underline{\beta}(\alpha), \bar{\beta}(\alpha)]$ and writing that as a convex combination of $\tilde{\tilde{\beta}}(\alpha) = 1$ (for which $\pi = PA$ is optimal by Proposition 1) and $\underline{\beta}(\alpha) - \epsilon$ will result in a contradiction. Thus, $\pi = PD$ is optimal at both extreme points $\tilde{\beta}(\alpha) = 0$ and $\underline{\beta}(\alpha) - \epsilon$. Hence, $\pi = PD$ is also optimal at their convex combination, $\beta(\alpha)$, by Lemma 1. If, on the other hand, (b) holds, write $\beta(\alpha)$ as a convex combination of $\overline{\beta}(\alpha) + \epsilon$ and $\tilde{\tilde{\beta}}(\alpha) = 1$. Then, similar to the discussion of part (a), notice that by definition of $\beta_2(\alpha)$, $\pi = PA$ is optimal at $\overline{\beta}(\alpha) + \epsilon$. Moreover, by Proposition 1, $\pi = PA$ is also optimal at $\tilde{\tilde{\beta}}(\alpha) = 1$. Thus, from Lemma 1 we see that $\pi = PA$ should be also optimal at $\beta(\alpha)$. This completes the proof. □

Proof of Proposition 3 - Part (i). We develop a *Markov Decision Process (MDP)* model to show the optimality in the expected sense. It should be noted that the underlying problem is in the class of *multi-armed restless bandit problems*, which are usually hard to analyze. Since beds are not limited (e.g., larger than the number of patients in the clearing model), suppose, without loss of generality, that at the beginning all patients are in state W_1, i.e., in the initial waiting state depicted in Figure 2.5. The ith waiting stage, W_i, is followed by a treatment stage, T_i. The duration of waiting stages and treatment stages are independent of each other and exponentially

distributed with rates denoted by γ and μ, respectively. Suppose the maximum number of interactions with the physician is denoted by $\bar{k}$, and $W_{\bar{k}+1}$ denotes the final nurse visit before disposition (i.e., stage FW in Figure 2.5). For the ease of notation, we also assume stage $T_{\bar{k}+1}$ represents the disposition stage. That is, we assume every patients who leaves the ED goes to (absorbing) stage $T_{\bar{k}+1}$. The LOS of a patient in our clearing model is then equal to the time that s/he leaves stage $W_{\bar{k}+1}$ to enter $T_{\bar{k}+1}$. Let p_k denote the probability that a patient who is in treatment stage k, T_k, is having its final treatment by the physician and will go to the final treatment by nurse, $W_{\bar{k}+1}$, afterwards. Assume p_k is increasing in k (that is being in a higher treatment stage is associated with a higher chance of being in the final treatment stage) and $p_{\bar{k}} = 1$. The state of the system then can be represented by $(\mathbf{X}, \mathbf{Y})$ with $\mathbf{X} = (x_1, x_2, \cdots, x_{\bar{k}+1})$ and $\mathbf{Y} = (y_1, y_2, \cdots, y_{\bar{k}+1})$, where x_i and y_i denote the number of patients in ith stage of treatment and wait (T_i and W_i), respectively. Let N denote the total number of patients at time 0. The goal is to dynamically control the location of the physician, denoted by l, to go from state $(N, 0, \cdots, 0)$ to state $(0, 0, \cdots, N)$ with the minimum expected average LOS or equivalently with the minimum sum of patient completion times. Now, using uniformization with rate $\psi = N\gamma + \mu < \infty$, we can consider the discrete time version of the problem (where the times between consecutive events are i.i.d and exponentially distributed with rate ψ). Doing so and denoting the optimal remaining cost when the system is at state $(\mathbf{X}, \mathbf{Y})$ with $J(\mathbf{X}, \mathbf{Y})$, we have the

following optimality equation (with the terminal condition $J(0,0,\cdots,N)=0$):

$$
\begin{aligned}
J(\mathbf{X},\mathbf{Y}) = \frac{1}{\psi}\Big[& \sum_{i=1}^{\bar{k}} x_i + \sum_{i=1}^{\bar{k}+1} y_i \\
& + \mu \min_{l\in\mathcal{L}(\mathbf{x})} \Big\{ \sum_{k=1}^{\bar{k}} \mathbb{1}\{l=k\}\big[p_k\, J(\mathbf{X}-e_k, \mathbf{Y}+e_{\bar{k}+1}) \\
& \qquad + (1-p_k)\, J(\mathbf{X}-e_k, \mathbf{Y}+e_{k+1})\big]\Big\} \\
& + \gamma \sum_{i=1}^{\bar{k}+1} y_i\, J(\mathbf{X}+e_i, \mathbf{Y}-e_i) \\
& + \big(\psi - \gamma \sum_{i=1}^{\bar{k}+1} y_i - \mu \mathbb{1}\big\{ \sum_{ki=1}^{\bar{k}} x_i \geq 1\big\}\big) J(\mathbf{X},\mathbf{Y})\Big],
\end{aligned}
\tag{2.3}
$$

where e_k is a row vector of size $\bar{k}+1$ with a one in kth element and zero everywhere else, and $\mathcal{L}(\mathbf{X}) = \{i \leq \bar{k} : x_i \geq 1\}$ is the set of possible locations to allocate the physician when $\mathbf{X}$ is the first part of the state. The first line in the above optimality equation represents the current cost (every patient's completion time who is still in the ED is delayed for one unit of uniformized time). The second line is the event related to treating a patient by the physician. The third line represents the event that a patient moves from a waiting stage to a treatment stage, and the fourth line represents the self-loop event. (Notice that since preemption is allowed, using a sample path argument, it can be easily shown that forced idling is suboptimal. Therefore, without loss of generality the term in the self-loop with coefficient μ is independent of the control action, l.) Also, a finite horizon version of the above MDP can be considered using the following optimality equation with terminal condition

$J_0(\mathbf{X}, \mathbf{Y}) = 0$ for every state $(\mathbf{X}, \mathbf{Y})$ and $n \in \mathbb{N}$:

$$
\begin{aligned}
J_{n+1}(\mathbf{X}, \mathbf{Y}) = \frac{1}{\psi} \Big[& \sum_{i=1}^{\bar{k}} x_i + \sum_{i=1}^{\bar{k}+1} y_i \\
& + \mu \min_{l \in \mathcal{L}(\mathbf{x})} \Big\{ \sum_{k=1}^{\bar{k}} \mathbb{1}\{l = k\} \big[p_k \, J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) \\
& \qquad + (1 - p_k) \, J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}) \big] \Big\} \\
& + \gamma \sum_{i=1}^{\bar{k}+1} y_i \, J_n(\mathbf{X} + e_i, \mathbf{Y} - e_i) \\
& + \big(\psi - \gamma \sum_{i=1}^{\bar{k}+1} y_i - \mu \mathbb{1}\big\{ \sum_{i=1}^{\bar{k}} x_i \geq 1 \big\} \big) J_n(\mathbf{X}, \mathbf{Y}) \Big], \qquad (2.4)
\end{aligned}
$$

where $J_n(\mathbf{X}, \mathbf{Y})$ denotes the optimal remaining cost when the state is $(\mathbf{X}, \mathbf{Y})$ and there are n periods to go. (Notice that $J_n(\mathbf{X}, \mathbf{Y}) \to J(\mathbf{X}, \mathbf{Y})$ as $n \to \infty$ since there is an absorbing state.) To show that the PO policy which prescribes serving the "old" patient in the most downstream stage is optimal, we use induction on n. First notice that for $n = 1$ all policies are the same considering the minimization in (2.4), since $J_0(\mathbf{X}, \mathbf{Y}) = 0$ for every state $(\mathbf{X}, \mathbf{Y})$. Now, suppose it is optimal to follow PO policy at any state when in period n. We show that it is optimal to follow PO at any state in period $n+1$ as well. To this end, consider period $n+1$ and an arbitrary state $(\mathbf{X}, \mathbf{Y})$. Suppose in state $(\mathbf{X}, \mathbf{Y})$ treatment stage k^* is the the most downstream stage with an available patient. To show that allocating the physician to stage $1 \leq k^* \leq \bar{k}$ is optimal in $n+1$, suppose there is also another stage $k < k^*$ with an available patient at state $(\mathbf{X}, \mathbf{Y})$ (i.e., with $x_k \geq 1$ and $x_{k^*} \geq 1$). Then considering the minimization in (2.4), to show that serving stage k^* in period $n+1$ is optimal, it is sufficient to

show that for any such k, we have:

$$\underline{\textbf{Property i:}} \qquad \begin{aligned} & p_{k^*}\, J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_{k^*})\, J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{k^*+1}) \\ & \leq p_k\, J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{\bar{k}+1}) + (1 - p_k)\, J_n(\mathbf{X} - e_k, \mathbf{Y} + e_{k+1}). \end{aligned} \tag{2.5}$$

We show the above property of the optimal cost function along with the following property:

$$\underline{\textbf{Property ii:}} \qquad \begin{aligned} & p_k^*\, J_n(\mathbf{X} + e_{\bar{k}+1} - e_{k^*}, \mathbf{Y}) + (1 - p_k^*)\, J_n(\mathbf{X} + e_{k^*+1} - e_{k*}, \mathbf{Y}) \\ & p_k\, J_n(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y}) + (1 - p_k)\, J_n(\mathbf{X} + e_{k+1} - e_k, \mathbf{Y}). \end{aligned} \tag{2.6}$$

In other words, we assume Properties i and ii hold for $n-1$, and show that they both hold for n as well. First, we show Property i. To do so, we build an upper bound for the LHS of (2.5) using suboptimal actions and show that this upper bound is less than the RHS of this inequality. The upper bound for the LHS can be obtained by suboptimally allocating the physician to treatment stage k in period n and then following the optimal policy (i.e., PO) in the remaining periods. To this end, consider state $(\mathbf{X}-e_{k^*}, \mathbf{Y}+e_{\bar{k}+1})$ in period n and use the suboptimal but feasible (since $x_k \geq 1$) action $l = k$ to obtain an upper bound for $J_n(\mathbf{X} - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1})$. Doing so we have:

$$
\begin{aligned}
J_n(\mathbf{X}-e_{k^*},\mathbf{Y}+e_{\bar{k}+1}) \le \frac{1}{\psi}\Big[&\big(\sum_{i=1}^{\bar{k}} x_i - 1\big) + \big(\sum_{i=1}^{\bar{k}+1} y_i + 1\big) \\
&+ \mu\Big[p_k\, J_{n-1}(\mathbf{X}-e_k-e_{k^*},\mathbf{Y}+e_{\bar{k}+1}+e_{\bar{k}+1}) \\
&\qquad + (1-p_k)\, J_{n-1}(\mathbf{X}-e_k-e_{k^*},\mathbf{Y}+e_{k+1}+e_{\bar{k}+1})\Big] \\
&+ \gamma \sum_{i=1}^{\bar{k}+1} y_i\, J_{n-1}(\mathbf{X}+e_i-e_{k^*},\mathbf{Y}-e_i+e_{\bar{k}+1}) \\
&+ \gamma J_{n-1}(\mathbf{X}+e_{\bar{k}+1}-e_{k^*},\mathbf{Y}) \\
&+ \big(\psi - \gamma(\sum_{i=1}^{\bar{k}+1} y_i + 1) \\
&\qquad - \mu \mathbb{1}\big\{\sum_{i=1}^{\bar{k}} x_i \ge 1\big\}\big) J_{n-1}(\mathbf{X}-e_{k^*},\mathbf{Y}+e_{\bar{k}+1})\Big].
\end{aligned}
\tag{2.7}
$$

Similarly, using the suboptimal but feasible action $l = k$ at state $(\mathbf{X}-e_{k^*},\mathbf{Y}+e_{k^*+1})$, we obtain an upper bound for $J_n(\mathbf{X}-e_{k^*},\mathbf{Y}+e_{k^*+1})$:

$$
\begin{aligned}
J_n(\mathbf{X}-e_{k^*},\mathbf{Y}+e_{k^*+1}) \leq \frac{1}{\psi}\Big[&\big(\sum_{i=1}^{\bar{k}} x_i - 1\big) + \big(\sum_{i=1}^{\bar{k}+1} y_i + 1\big) \\
&+ \mu\left[p_k\, J_{n-1}(\mathbf{X}-e_k-e_{k^*},\mathbf{Y}+e_{\bar{k}+1}+e_{k^*+1})\right. \\
&\qquad \left.+ (1-p_k)\, J_{n-1}(\mathbf{X}-e_k-e_{k^*},\mathbf{Y}+e_{k+1}+e_{k^*+1})\right] \\
&+ \gamma \sum_{i=1}^{\bar{k}+1} y_i\, J_{n-1}(\mathbf{X}+e_i-e_{k^*},\mathbf{Y}-e_i+e_{k*+1}) \\
&+ \gamma J_{n-1}(\mathbf{X}+e_{k^*+1}-e_{k^*},\mathbf{Y}) \\
&+ \big(\psi - \gamma(\sum_{i=1}^{\bar{k}+1} y_i + 1) \\
&\qquad - \mu \mathbb{1}\big\{\sum_{i=1}^{\bar{k}} x_i \geq 1\big\}\big) J_{n-1}(\mathbf{X}-e_{k^*},\mathbf{Y}+e_{k^*+1})\Big].
\end{aligned}
\tag{2.8}
$$

Now multiplying both sides of (2.7) by p_{k^*}, both sides of (2.8) by $(1-p_{k^*})$, and summing up the results we have:

$$
\begin{aligned}
&p_{k^*}\, J_n(\mathbf{X}-e_{k^*},\mathbf{Y}+e_{\bar{k}+1}) + (1-p_{k^*})\, J_n(\mathbf{X}-e_{k^*},\mathbf{Y}+e_{k^*+1}) \leq \\
&\frac{1}{\psi}\Big[\big(\sum_{i=1}^{\bar{k}} x_i - 1\big) + \big(\sum_{i=1}^{\bar{k}+1} y_i + 1\big) + \mu\,\big[p_{k^*}\, p_k\, J_{n-1}(\mathbf{X}-e_k-e_{k^*},\mathbf{Y}+e_{\bar{k}+1}+e_{\bar{k}+1}) \\
&\qquad\qquad + p_{k^*}\,(1-p_k)\, J_{n-1}(\mathbf{X}-e_k-e_{k^*},\mathbf{Y}+e_{k+1}+e_{\bar{k}+1}) \\
&\qquad\qquad + (1-p_{k^*})\, p_k\, J_{n-1}(\mathbf{X}-e_k-e_{k^*},\mathbf{Y}+e_{k^*+1}+e_{\bar{k}+1}) \\
&\qquad\qquad + (1-p_{k^*})\,(1-p_k)\, J_{n-1}(\mathbf{X}-e_k-e_{k^*},\mathbf{Y}+e_{k^*+1}+e_{k+1})\big] \\
&\quad + \gamma \sum_{i=1}^{\bar{k}+1} y_i\,\big[p_{k^*}\, J_{n-1}(\mathbf{X}+e_i-e_{k^*},\mathbf{Y}-e_i+e_{\bar{k}+1}) + \\
&\qquad\qquad (1-p_{k^*}) J_{n-1}(\mathbf{X}+e_i-e_{k^*},\mathbf{Y}-e_i+e_{k^*+1})\big] \\
&\quad + \gamma\,[p_{k^*}\, J_{n-1}(\mathbf{X}+e_{\bar{k}+1}-e_{k^*},\mathbf{Y}) + \,(1-p_{k^*})\, J_{n-1}(\mathbf{X}+e_{k^*+1}-e_{k^*},\mathbf{Y})] \\
&\quad + \bar{\psi}\,\big(\,p_{k^*}\, J_{n-1}(\mathbf{X}-e_{k^*},\mathbf{Y}+e_{\bar{k}+1}) + (1-p_{k^*})\, J_{n-1}(\mathbf{X}-e_{k^*},\mathbf{Y}+e_{k^*+1})\big)\Big], \quad (2.9)
\end{aligned}
$$

where, for the ease of notation, we let $\bar{\psi}$ denote the self-loop rate, i.e., $\bar{\psi} = \big(\psi - \gamma(\sum_{i=1}^{\bar{k}+1} y_i + 1) - \mu \mathbb{1}\big\{\sum_{i=1}^{\bar{k}} x_i \geq 1\big\}\big)$. Now in the above upper bound, using the induction hypothesis, we can replace the terms with coefficient γ to obtain another upper bound. Using Property i and ii for the first and second terms with coefficient γ, we have:

$$
\begin{aligned}
&p_{k^*}\, J_n(\mathbf{X}-e_{k^*},\mathbf{Y}+e_{\bar{k}+1}) + (1-p_{k^*})\, J_n(\mathbf{X}-e_{k^*},\mathbf{Y}+e_{k^*+1}) \leq \\
&\frac{1}{\psi}\Big[\big(\sum_{i=1}^{\bar{k}} x_i - 1\big) + \big(\sum_{i=1}^{\bar{k}+1} y_i + 1\big) + \mu\,\big[p_{k^*}\, p_k\, J_{n-1}(\mathbf{X}-e_k-e_{k^*},\mathbf{Y}+e_{\bar{k}+1}+e_{\bar{k}+1}) \\
&\qquad\qquad + p_{k^*}\,(1-p_k)\, J_{n-1}(\mathbf{X}-e_k-e_{k^*},\mathbf{Y}+e_{k+1}+e_{\bar{k}+1}) \\
&\qquad\qquad + (1-p_{k^*})\, p_k\, J_{n-1}(\mathbf{X}-e_k-e_{k^*},\mathbf{Y}+e_{k^*+1}+e_{\bar{k}+1}) \\
&\qquad\qquad + (1-p_{k^*})\,(1-p_k)\, J_{n-1}(\mathbf{X}-e_k-e_{k^*},\mathbf{Y}+e_{k^*+1}+e_{k+1})\big] \\
&\quad + \gamma \sum_{i=1}^{\bar{k}+1} y_i\big[\, p_k\, J_{n-1}(\mathbf{X}+e_i-e_k,\mathbf{Y}-e_i+e_{\bar{k}+1}) \\
&\qquad\quad + (1-p_k)\, J_{n-1}(\mathbf{X}+e_i-e_k,\mathbf{Y}-e_i+e_{k+1})\big] \\
&\quad + \gamma\,[p_k\, J_{n-1}(\mathbf{X}+e_{\bar{k}+1}-e_k,\mathbf{Y}) + (1-p_k)\, J_{n-1}(\mathbf{X}+e_{k+1}-e_k,\mathbf{Y})] \\
&\quad + \bar{\psi}\,\big(p_k\, J_{n-1}(\mathbf{X}-e_k,\mathbf{Y}+e_{\bar{k}+1}) + (1-p_k)\, J_{n-1}(\mathbf{X}-e_k,\mathbf{Y}+e_{k+1})\big)\Big]. \qquad (2.10)
\end{aligned}
$$

Thus, we have obtained an upper bound for the LHS of (2.5). Now consider the RHS of (2.5) and first for state $(\mathbf{X}-e_k,\mathbf{Y}+e_{\bar{k}+1})$ use (2.4) to obtain $J_n(\mathbf{X}-e_k,\mathbf{Y}+e_{\bar{k}+1})$. Note that, by the induction hypothesis, PO is optimal in period n. Hence, it is optimal to assign the physician to treatment stage k^* in period n at state $(\mathbf{X}-e_k,\mathbf{Y}+e_{\bar{k}+1})$, since k^* is the most down-stream treatment stage with an available patient when state is $(\mathbf{X},\mathbf{Y})$ (and hence when state is $(\mathbf{X}-e_k,\mathbf{Y}+e_{\bar{k}+1})$). Thus,

using (2.4) we have:

$$
\begin{aligned}
J_n(\mathbf{X}-e_k,\mathbf{Y}+e_{\bar{k}+1}) = \frac{1}{\psi}\Big[&\big(\sum_{i=1}^{\bar{k}} x_i - 1\big) + \big(\sum_{i=1}^{\bar{k}+1} y_i + 1\big) \\
&+ \mu\left[p_{k^*}\, J_{n-1}(\mathbf{X}-e_{k^*}-e_k,\mathbf{Y}+e_{\bar{k}+1}+e_{\bar{k}+1})\right. \\
&\qquad \left. + (1-p_{k^*})\, J_{n-1}(\mathbf{X}-e_{k^*}-e_k,\mathbf{Y}+e_{k^*+1}+e_{\bar{k}+1})\right] \\
&+ \gamma \sum_{i=1}^{\bar{k}+1} y_i\, J_{n-1}(\mathbf{X}+e_i-e_k,\mathbf{Y}-e_i+e_{\bar{k}+1}) \\
&+ \gamma J_{n-1}(\mathbf{X}+e_{\bar{k}+1}-e_k,\mathbf{Y}) \\
&+ \big(\psi - \gamma(\sum_{i=1}^{\bar{k}+1} y_i + 1) \\
&\qquad - \mu \mathbb{1}\big\{\sum_{i=1}^{\bar{k}} x_i \geq 1\big\}\big) J_{n-1}(\mathbf{X}-e_k,\mathbf{Y}+e_{\bar{k}+1})\Big].
\end{aligned}
\tag{2.11}
$$

Similarly, using (2.4) to obtain $J_n(\mathbf{X}-e_k,\mathbf{Y}+e_{k+1})$ we have:

$$
\begin{aligned}
J_n(\mathbf{X}-e_k,\mathbf{Y}+e_{k+1}) = \frac{1}{\psi}\Big[&\big(\sum_{i=1}^{\bar{k}} x_i - 1\big) + \big(\sum_{i=1}^{\bar{k}+1} y_i + 1\big) \\
&+ \mu\left[p_{k^*}\, J_{n-1}(\mathbf{X}-e_{k^*}-e_k,\mathbf{Y}+e_{k+1}+e_{\bar{k}+1})\right. \\
&\qquad \left. + (1-p_{k^*})\, J_{n-1}(\mathbf{X}-e_{k^*}-e_k,\mathbf{Y}+e_{k^*+1}+e_{k+1})\right] \\
&+ \gamma \sum_{i=1}^{\bar{k}+1} y_i\, J_{n-1}(\mathbf{X}+e_i-e_k,\mathbf{Y}-e_i+e_{k+1}) \\
&+ \gamma J_{n-1}(\mathbf{X}+e_{k+1}-e_k,\mathbf{Y}) \\
&+ \big(\psi - \gamma(\sum_{i=1}^{\bar{k}+1} y_i + 1) \\
&\qquad - \mu \mathbb{1}\big\{\sum_{i=1}^{\bar{k}} x_i \geq 1\big\}\big) J_{n-1}(\mathbf{X}-e_k,\mathbf{Y}+e_{k+1})\Big].
\end{aligned}
\tag{2.12}
$$

Now multiplying both sides of (2.11) by p_k, both sides of (2.12) by $(1-p_k)$, and summing up the results we have:

$$
\begin{aligned}
&p_k\, J_n(\mathbf{X}-e_k, \mathbf{Y}+e_{\bar{k}+1}) + (1-p_k)\, J_n(\mathbf{X}-e_k, \mathbf{Y}+e_{k+1}) = \\
&\frac{1}{\psi}\Big[\big(\sum_{i=1}^{\bar{k}} x_i - 1\big) + \big(\sum_{i=1}^{\bar{k}+1} y_i + 1\big) + \mu\,\big[p_{k^*}\, p_k\, J_{n-1}(\mathbf{X}-e_k-e_{k^*}, \mathbf{Y}+e_{\bar{k}+1}+e_{\bar{k}+1}) \\
&\qquad\qquad + p_{k^*}\,(1-p_k)\, J_{n-1}(\mathbf{X}-e_k-e_{k^*}, \mathbf{Y}+e_{k+1}+e_{\bar{k}+1}) \\
&\qquad\qquad + (1-p_{k^*})\, p_k\, J_{n-1}(\mathbf{X}-e_k-e_{k^*}, \mathbf{Y}+e_{k^*+1}+e_{\bar{k}+1}) \\
&\qquad\qquad + (1-p_{k^*})\,(1-p_k)\, J_{n-1}(\mathbf{X}-e_k-e_{k^*}, \mathbf{Y}+e_{k^*+1}+e_{k+1})\big] \\
&\quad + \gamma \sum_{i=1}^{\bar{k}+1} y_i \big[\, p_k\, J_{n-1}(\mathbf{X}+e_i-e_k, \mathbf{Y}-e_i+e_{\bar{k}+1}) \\
&\qquad + (1-p_k)\, J_{n-1}(\mathbf{X}+e_i-e_k, \mathbf{Y}-e_i+e_{k+1})\big] \\
&\quad + \gamma\,[p_k\, J_{n-1}(\mathbf{X}+e_{\bar{k}+1}-e_k, \mathbf{Y}) + (1-p_k)\, J_{n-1}(\mathbf{X}+e_{k+1}-e_k, \mathbf{Y})] \\
&\quad + \bar{\psi}\,\big(p_k\, J_{n-1}(\mathbf{X}-e_k, \mathbf{Y}+e_{\bar{k}+1}) + (1-p_k)\, J_{n-1}(\mathbf{X}-e_k, \mathbf{Y}+e_{k+1})\big)\Big], \qquad (2.13)
\end{aligned}
$$

where, for the ease of notation, we again let $\bar{\psi} = \big(\psi - \gamma(\sum_{i=1}^{\bar{k}+1} y_i + 1) - \mu \mathbb{1}\big\{\sum_{i=1}^{\bar{k}} x_i \geq 1\big\}\big)$. Notice that RHS of (2.13) is equal to the upper bound of the LHS of (2.5) derived in (2.10). Thus, Property i holds for every n by induction, and hence the PO is optimal in every period.

To complete the proof, it remains to show Property ii. To do so, we use the same technique used to show Property i. First, notice that for $n = 0$ (or $n = 1$) this property is trivial. Next suppose it holds for $n-1$. To show that it would also hold for n, we use suboptimal actions to obtain an upper bound for the LHS of (2.6) and show that this upper bound is equal to its RHS. To do so, consider states $(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y})$ and $(\mathbf{X} + e_{k^*+1} - e_k, \mathbf{Y})$, and for each one, to obtain an upper bound, use the optimality equation (2.4) but with suboptimal actions $l = k$. Then multiply the upper bound obtained for $J(\mathbf{X}+e_{\bar{k}+1}-e_k*, \mathbf{Y})$ and $J(\mathbf{X}+e_{k^*+1}-e_k^*, \mathbf{Y})$

by p_{k^*} and $1-p_{k^*}$, respectively. Summing up the results, we gain the following upper bound for the LHS of (2.5):

$$
\begin{aligned}
& p_{k^*}\, J_n(\mathbf{X}+e_{\bar{k}+1}-e_{k^*},\mathbf{Y}) + \,(1-p_{k^*})\, J_n(\mathbf{X}+e_{k^*+1}-e_{k*},\mathbf{Y}) \leq \\
& \frac{1}{\psi}\Big[-p_{k^*} + \big(\sum_{i=1}^{\bar{k}} x_i\big) + \big(\sum_{i=1}^{\bar{k}+1} y_i\big) \\
& \quad + \mu\,\big[p_{k^*}\, p_k\, J_{n-1}(\mathbf{X}+e_{\bar{k}+1}-e_k-e_{k^*},\mathbf{Y}+e_{\bar{k}+1}) \\
& \qquad + p_{k^*}\,(1-p_k)\, J_{n-1}(\mathbf{X}+e_{\bar{k}+1}-e_k-e_{k^*},\mathbf{Y}+e_{k+1}) \\
& \qquad + (1-p_{k^*})\, p_k\, J_{n-1}(\mathbf{X}+e_{k^*+1}-e_k-e_{k^*},\mathbf{Y}+e_{\bar{k}+1}) \\
& \qquad + (1-p_{k^*})\,(1-p_k)\, J_{n-1}(\mathbf{X}+e_{k^*+1}-e_k-e_{k^*},\mathbf{Y}+e_{k+1})\big] \\
& \quad + \gamma \sum_{i=1}^{\bar{k}+1} y_i\big[\, p_{k^*}\, J_{n-1}(\mathbf{X}+e_{\bar{k}+1}+e_i-e_{k^*},\mathbf{Y}-e_i) \\
& \qquad\qquad + (1-p_{k^*})\, J_{n-1}(\mathbf{X}+e_{k^*+1}+e_i-e_{k^*},\mathbf{Y}-e_i)\big] \\
& \quad + \psi\,\big(p_k\, J_{n-1}(\mathbf{X}+e_{\bar{k}+1}-e_{k^*},\mathbf{Y}) + (1-p_k)\, J_{n-1}(\mathbf{X}+e_{k^*+1}-e_{k*},\mathbf{Y})\big)\Big]. \quad (2.14)
\end{aligned}
$$

Now, using the optimality equation (2.4) to derive $J_n(\mathbf{X}+e_{\bar{k}+1}-\mathbf{e}_k,\mathbf{Y})$ and $J_n(\mathbf{X}+e_{k+1}-\mathbf{e}_k,\mathbf{Y})$, and then multiplying them by p_k and $1-p_k$, respectively, and finally summing up the results we get the following equality for the RHS of (2.5). (Notice that by the induction hypothesis assigning the physician to k^* is optimal when computing

$J_n(\mathbf{X} + e_{\bar{k}+1} - \mathbf{e}_k, \mathbf{Y})$ and $J_n(\mathbf{X} + e_{k+1} - \mathbf{e}_k, \mathbf{Y})$.)

$$
\begin{aligned}
& p_k\, J_n(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y}) + (1 - p_k)\, J_n(\mathbf{X} + e_{k+1} - e_k, \mathbf{Y}) = \\
& \frac{1}{\psi} \Big[- p_k + \big(\sum_{i=1}^{\bar{k}} x_i\big) + \big(\sum_{i=1}^{\bar{k}+1} y_i\big) \\
& \quad + \mu \big[p_{k^*}\, p_k\, J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) \\
& \qquad + p_{k^*}\, (1 - p_k)\, J_{n-1}(\mathbf{X} + e_{k+1} - e_k - e_{k^*}, \mathbf{Y} + e_{\bar{k}+1}) \\
& \qquad + (1 - p_{k^*})\, p_k\, J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k - e_{k^*}, \mathbf{Y} + e_{k^*+1}) \\
& \qquad + (1 - p_{k^*})\, (1 - p_k)\, J_{n-1}(\mathbf{X} + e_{k+1} - e_k - e_{k^*}, \mathbf{Y} + e_{k^*+1}) \big] \\
& \quad + \gamma \sum_{i=1}^{\bar{k}+1} y_i \big[p_k\, J_{n-1}(\mathbf{X} + e_{\bar{k}+1} + e_i - e_k, \mathbf{Y} - e_i) \\
& \qquad\qquad + (1 - p_k)\, J_{n-1}(\mathbf{X} + e_{k+1} + e_i - e_k, \mathbf{Y} - e_i) \big] \\
& \quad + \psi \big(p_k\, J_{n-1}(\mathbf{X} + e_{\bar{k}+1} - e_k, \mathbf{Y}) + (1 - p_k)\, J_{n-1}(\mathbf{X} + e_{k+1} - e_k, \mathbf{Y}) \big) \Big]. \qquad (2.15)
\end{aligned}
$$

Now, notice that since $k^* > k$, by assumption we have $p_{k^*} \geq p_k$. Next, using the induction hypothesis and since $p_{k^*} \geq p_k$, it is easy to show that the upper bound obtained in (2.14) is less than or equal to (2.15), which establishes Property ii for n and completes the proof. □

Proof of Proposition 3 - Part (ii). We use a sample path argument to show the result in the almost sure sense. Consider the probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and similar to the proof of part (i), without loss of generality, suppose at time 0, all of the N patients in the clearing model are in state W_1, i.e., in the initial waiting state depicted in Figure 2.5. Let $w_1^n(\omega)$ be the realized duration of the initial waiting stage, W_1, for patient $n \in \{1, \cdots, N\}$ under sample path $\omega \in \Omega$. Let G be the set of all admissible (Markovovian or non-Markovian) policies and $TTFT^{g,n}(\omega)$ be the Time To First Treatment of patient n under policy $g \in G$ and sample path $\omega \in \Omega$. Notice that $TTFT^{g,n}(\omega) \geq w_1^n(\omega)$ for every $g \in G$, every $\omega \in \Omega$, and every $n \in \{1, \cdots, N\}$, since a patient cannot been seen before s/he finishes stage W_1.

Therefore, $\inf_{g\in G} TTFT^{g,n}(\omega) \geq w_1^n(\omega)$. Now notice that for the underlying Prioritize New (PN) policy, which instructs the physician to initialize a new patient journey whenever possible (perhaps by preempting other tasks), $TTFT^{PN,n}(\omega) = w_1^n(\omega)$ (for every $\omega \in \Omega$, and every $n \in \{1, \cdots, N\}$). Thus, the PN obtains the minimum TTFT of every patient along every sample path. Therefore, PN also minimizes the average TTFT of patients with probability one (i.e., in the almost sure sense). □

2.9 Appendix B: Computations Under Imperfect Classification

Assume $I \in \{A, D\}$ represents the true identity of the patient (Admit or Discharge) and $\omega \in \{A, D\}$ is the signaled/identified class. Let $\gamma_A = Pr(\omega = D|I = A)$ and $\gamma_D = Pr(\omega = A|I = D)$. Next, if $\tilde{\gamma}_A = Pr(I = A|\omega = D)$ and $\tilde{\gamma}_D = Pr(I = D|\omega = A)$ represent the misclassification probabilities, with $\alpha = Pr(I = A)$, using Bayes rule we have:

$$\begin{aligned}
\tilde{\gamma}_A &= Pr(I = A|\omega = D) = \frac{\alpha\,\gamma_A}{\alpha\gamma_A + (1-\alpha)(1-\gamma_D)}, \\
\tilde{\gamma}_D &= Pr(I = D|\omega = A) = \frac{(1-\alpha)\,\gamma_D}{\alpha(1-\gamma_A) + (1-\alpha)\gamma_D}.
\end{aligned}$$

To isolate the effect of misclassification errors, we eliminate variability in the treatment times, X_A and X_D so that $Pr(X_A = \mu_A) = 1$ and $Pr(X_D = \mu_D) = 1$. Moreover, for the ease of computations, we consider a collaborative service environment whenever the system is working in the pooling mode (i.e., under pooling or under streaming after one stream runs out of patients). Collaborative assumption means that the two servers work together on one patient at a time with service times of $\mu_A/2$ for admits and $\mu_D/2$ for discharges.

Let n be the total number of patients in the clearing system. Suppose N_A and $N_D = n - N_A$ denote the random variable representing the number of patients that

are identified as A and D, respectively. Let $\tilde{N}_A$ and $\tilde{N}_D$ be the random variables representing last patients of type A and D that are seen before the system moves to a pooling scenario, reactively. Next notice that given N_A (and hence $N_D = n - N_A$), $\tilde{N}_A$ and $\tilde{N}_D$, expected TTFT of Admits under Streaming can be computed by:

$$E\big[TTFT_A^S | N_A = n_A, \tilde{N}_A = \tilde{n}_A, \tilde{N}_D = \tilde{n}_D\big] = \frac{1}{(1-\tilde{\gamma}_D)n_A + \tilde{\gamma}_A(n-n_A)} \times$$

$$\Big[(1-\tilde{\gamma}_D)\Big[\sum_{j=1}^{\tilde{n}_A}\sum_{k=0}^{j-1}\binom{j-1}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{j-k-1}\big(k\mu_D + (j-k-1)\mu_A\big)$$
$$+ \sum_{j=\tilde{n}_A+1}^{n_A}\Big[\sum_{k=0}^{\tilde{n}_A}\binom{\tilde{n}_A}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{\tilde{n}_A-k}\big(k\mu_D + (\tilde{n}_A-k)\mu_A\big)$$
$$+ \sum_{k=0}^{j-\tilde{n}_A-1}\binom{j-\tilde{n}_A-1}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{j-\tilde{n}_A-1-k}\big(k\frac{\mu_D}{2}$$
$$+ (j-\tilde{n}_A-1-k)\frac{\mu_A}{2}\big)\Big]\Big]$$
$$+\tilde{\gamma}_A\Big[\sum_{j=1}^{\tilde{n}_D}\sum_{k=0}^{j-1}\binom{j-1}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{j-k-1}\big(k\mu_A + (j-k-1)\mu_D\big)$$
$$+ \sum_{j=\tilde{n}_D+1}^{n-n_A}\Big[\sum_{k=0}^{\tilde{n}_D}\binom{\tilde{n}_D}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{\tilde{n}_D-k}\big(k\mu_A + (\tilde{n}_D-k)\mu_D\big)$$
$$+ \sum_{k=0}^{j-\tilde{n}_D-1}\binom{j-\tilde{n}_D-1}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{j-\tilde{n}_D-1-k}\big(k\frac{\mu_A}{2}$$
$$+ (j-\tilde{n}_D-1-k)\frac{\mu_D}{2}\big)\Big]\Big]\Big]. \tag{2.16}$$

The first line in the above equation is the reciprocal of the number of A patients (either classified as A or D). The second line considers the jth patient in the stream of the patients classified as A and seen before the system moves to a pooling scenario (i.e., up to $\tilde{n}_A$) and computes its TTFT by conditioning on the number of D patients in front him. Similarly, the third and fourth line consider the jth patient in the

stream of the patients classified/signaled as A and seen after the system moves to a pooling scenario (i.e., after $\tilde{n}_A$). The second, third, and fourth lines are multiplied by $(1-\tilde{\gamma}_D)$ (i.e., the probability that a patient classified as A is truly A type) to give the total sum of TTFT of A patients who are also classified as A. Similarly, the fifth, sixth, and seventh lines compute the sum of TTFT of A patients who are classified as D.

Now if $g(n_A, \tilde{n}_A, \tilde{n}_D)$ represents the joint pdf of random variables $N_A, \tilde{N}_A, \tilde{N}_D$ then we have:

$$\begin{aligned}
\overline{TTFT}_A^S &= E\big[E[TTFT_A^S | N_A, \tilde{N}_A, \tilde{N}_D]\big] && (2.17)\\
&= \sum_{n_A=0}^{n}\sum_{\tilde{n}_A=0}^{n}\sum_{\tilde{n}_D=0}^{n} E[TTFT_A^S | N_A, \tilde{N}_A, \tilde{N}_D] g(n_A, \tilde{n}_A, \tilde{n}_D), && (2.18)
\end{aligned}$$

where $E[TTFT_A^S | N_A, \tilde{N}_A, \tilde{N}_D]$ is computed in (2.16). To compute $\overline{TTFT}_A^S$ using the above equation, it remains to derive $g(n_A, \tilde{n}_A, \tilde{n}_D)$. To derive $g(n_A, \tilde{n}_A, \tilde{n}_D)$ notice that:

$$\begin{aligned}
&g(n_A, \tilde{n}_A, \tilde{n}_D) = \\
&Pr(N_A = n_A, \tilde{N}_A = \tilde{n}_A, \tilde{N}_D = \tilde{n}_D) = \\
&Pr(\tilde{N}_A = \tilde{n}_A, \tilde{N}_D = \tilde{n}_D | N_A = n_A) \times Pr(N_A = n_A) = \\
&Pr(N_A = n_A)\big[Pr(\tilde{N}_A = \tilde{n}_A = N_A = n_A, \tilde{N}_D = \tilde{n}_D)\mathbb{1}\{\tilde{n}_D < n - n_A = n_D\} && (2.19)\\
&\qquad + Pr(\tilde{N}_A = \tilde{n}_A, \tilde{N}_D = \tilde{n}_D = n - N_A = n - n_A)\mathbb{1}\{\tilde{n}_A < n_A\} && (2.20)\\
&\qquad + Pr(\tilde{N}_A = \tilde{n}_A = N_A = n_A, \tilde{N}_D = \tilde{n}_D = n - N_A = n - n_A = n_D) \\
&\qquad\qquad \times \mathbb{1}\{\tilde{n}_A = n_A, \tilde{n}_D = n - n_A\}\big] && (2.21)
\end{aligned}$$

In above, Eq's (2.19), (2.20), and (2.21) correspond to the cases where the D stream is finished first, the A stream is finished first, and the case where one stream is done when the system is working on the last patient of the other stream, respectively.

Next notice that with $p = (1-\gamma_A)\alpha + \gamma_D(1-\alpha)$ denoting the probability that a patient is identified as A:

$$Pr(N_A = n_A) = \binom{n}{n_A} p^{n_A}(1-p)^{n-n_A}. \tag{2.22}$$

Let K_j^A and K_j^D be the random variables denoting the number of D type patients up to (and including) the jth patient in A and D streams, respectively. Then to compute (2.19), we need to compute the probability that the time required to see n_A patients in the A stream is between the time required to see $\tilde{n}_D$ and $\tilde{n}_D + 1$ patients in the D stream (so that $\tilde{n}_D$ is the last patient seen in the D stream before the system moves to the pooling scenario). we have:

$$\begin{aligned}
& Pr(\tilde{N}_A = \tilde{n}_A = N_A = n_A, \tilde{N}_D = \tilde{n}_D) \\
= \;& Pr\big((\tilde{n}_D - K^D_{\tilde{n}_D})\mu_A + K^D_{\tilde{n}_D}\mu_D \le (n_A - K^A_{n_A})\mu_A + K^A_{n_A}\mu_D\big) \\
- \;& Pr\big((n_A - K^A_{n_A})\mu_A + K^A_{n_A}\mu_D \ge (\tilde{n}_D + 1 - K^D_{\tilde{n}_D+1})\mu_A + K^D_{\tilde{n}_D+1}\mu_D\big) \\
= \;& Pr\big(K^A_{n_A} - K^D_{\tilde{n}_D} \le \mu_A \frac{n_A - \tilde{n}_D}{\mu_A - \mu_D}\big) \\
- \;& Pr\big(K^A_{n_A} - K^D_{\tilde{n}_D+1} \le \mu_A \frac{n_A - (\tilde{n}_D + 1)}{\mu_A - \mu_D}\big) \\
= \;& F_1\big(\mu_A \frac{n_A - \tilde{n}_D}{\mu_A - \mu_D}\big) - F_2\big(\mu_A \frac{n_A - (\tilde{n}_D + 1)}{\mu_A - \mu_D}\big)
\end{aligned} \tag{2.23}$$

where $F_1(\cdot)$ and $F_2(\cdot)$ are the CDF of the random variables $Z_1 = K^A_{n_A} - K^D_{\tilde{n}_D}$ and $Z_2 = K^A_{n_A} - K^D_{\tilde{n}_D+1}$, respectively. Similarly, to compute (2.20), we have:

$$
\begin{aligned}
& Pr(\tilde{N}_A = \tilde{n}_A, \tilde{N}_D = \tilde{n}_D = n - N_A = n - n_A = n_D) \\
= \;& Pr\big((\tilde{n}_A - K^A_{\tilde{n}_A})\mu_A + K^A_{\tilde{n}_A}\mu_D \le (n_D - K^D_{n_D})\mu_A + K^D_{n_D}\mu_D\big) \\
- \;& Pr\big((n_D - K^D_{n_D})\mu_A + K^D_{n_D}\mu_D \ge (\tilde{n}_A + 1 - K^A_{\tilde{n}_A+1})\mu_A + K^A_{\tilde{n}_A+1}\mu_D\big) \\
= \;& Pr\big(K^D_{n_D} - K^A_{\tilde{n}_A} \le \mu_A \frac{n_D - \tilde{n}_A}{\mu_A - \mu_D}\big) \\
- \;& Pr\big(K^D_{n_D} - K^D_{\tilde{n}_A+1} \le \mu_A \frac{n_D - (\tilde{n}_A + 1)}{\mu_A - \mu_D}\big) \\
= \;& F_3\big(\mu_A \frac{n_D - \tilde{n}_A}{\mu_A - \mu_D}\big) - F_4\big(\mu_A \frac{n_D - (\tilde{n}_A + 1)}{\mu_A - \mu_D}\big) \qquad (2.24)
\end{aligned}
$$

where $F_3(\cdot)$ and $F_4(\cdot)$ are the CDF of the random variables $Z_3 = K^D_{n_D} - K^A_{\tilde{n}_A}$ and $Z_4 = K^D_{n_D} - K^D_{\tilde{n}_A+1}$, respectively.

Next, to compute (2.21), we need to compute the probability that one stream

finishes when the system is working on the last patient of the other stream:

$$
\begin{aligned}
& Pr(\tilde{N}_A = \tilde{n}_A = N_A = n_A, \tilde{N}_D = \tilde{n}_D = n - N_A = n - n_A = n_D) \\
= \;& Pr(T^A_{n_A-1} < T^D_{n_D} \leq T^A_{n_A}) + Pr(T^D_{n_D-1} < T^A_{n_A} < T^D_{n_D}) \qquad (2.25) \\
= \;& Pr(((n_A - 1) - K^A_{n_A-1})\mu_A + K^A_{n_A-1}\mu_D < (n_D - K^D_{n_D})\mu_A + K^D_{n_D}\mu_D \\
\leq \;& (n_A - K^A_{n_A})\mu_A + K^A_{n_A}\mu_D) \\
+ \;& Pr(((n_D - 1) - K^D_{n_D-1})\mu_A + K^D_{n_D-1}\mu_D < (n_A - K^A_{n_A})\mu_A + K^A_{n_A}\mu_D \\
< \;& (n_D - K^D_{n_D})\mu_A + K^D_{n_D}\mu_D) \\
= \;& Pr(K^D_{n_D} - K^A_{n_A-1} < \mu_A \frac{n_D - (n_A - 1)}{\mu_A - \mu_D}) + Pr\big(K^A_{n_A} - K^D_{n_D-1} < \mu_A \frac{n_A - (n_D - 1)}{\mu_A - \mu_D}\big) \\
- \;& Pr((n_D - K^D_{n_D})\mu_A + K^D_{n_D}\mu_D > (n_A - K^A_{n_A})\mu_A + K^A_{n_A}\mu_D) \\
- \;& Pr((n_D - K^D_{n_D})\mu_A + K^D_{n_D}\mu_D \leq (n_A - K^A_{n_A})\mu_A + K^A_{n_A}\mu_D) \\
= \;& Pr(K^D_{n_D} - K^A_{n_A-1} < \mu_A \frac{n_D - (n_A - 1)}{\mu_A - \mu_D}) \\
+ \;& Pr\big(K^A_{n_A} - K^D_{n_D-1} < \mu_A \frac{n_A - (n_D - 1)}{\mu_A - \mu_D}\big) - 1 \\
= \;& F_5(\mu_A \frac{n_D - (n_A - 1)}{\mu_A - \mu_D}) + F_6(\mu_A \frac{n_A - (n_D - 1)}{\mu_A - \mu_D}) - 1 \\
- \;& Pr(K^D_{n_D} - K^A_{n_A-1} = \mu_A \frac{n_D - (n_A - 1)}{\mu_A - \mu_D}) \\
- \;& Pr(K^A_{n_A} - K^D_{n_D-1} = \mu_A \frac{n_A - (n_D - 1)}{\mu_A - \mu_D})
\end{aligned}
$$

where T in (2.25) is used to show the finish times of corresponding jobs, and $F_5(\cdot)$ and $F_5(\cdot)$ are CDFs of random variables $Z_5 = K^D_{n_D} - K^A_{n_A-1}$ and $Z_6 = K^A_{n_A} - K^D_{n_D-1}$. Now notice that random variables Z_1,...,Z_6 are each the difference between two independent binomial random variables with known parameters. Thus, CDFs F_1,...,F_6 are known. Therefore, $g(n_A, \tilde{n}_A, \tilde{n}_D)$ can be computed. As a result, the metric $\overline{TTFT}^S_A$ is completely computed.

Next, in a similar way, we compute the metric $\overline{LOS}^S_D$ (i.e., Expected Length of

Stay of D patients under Streaming): $\overline{LOS}_D^S =$

$$E\big[E[LOS_D^S|N_A,\tilde{N}_A,\tilde{N}_D]\big] = \sum_{n_A=0}^{n}\sum_{\tilde{n}_A=0}^{N}\sum_{\tilde{n}_D=0}^{N} E[LOS_D^S|N_A,\tilde{N}_A,\tilde{N}_D]g(n_A,\tilde{n}_A,\tilde{n}_D) \tag{2.26}$$

where:

$$\big[LOS_D^S|N_A=n_A,\tilde{N}_A=\tilde{n}_A,\tilde{N}_D=\tilde{n}_D\big] = \frac{1}{\tilde{\gamma}_D n_A+(1-\tilde{\gamma}_A)(n-n_A)}\times$$

$$\begin{aligned}
&\Big[\tilde{\gamma}_D\Big[\sum_{j=1}^{\tilde{n}_A}\sum_{k=0}^{j-1}\binom{j-1}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{j-k-1}\big((k+1)\mu_D+(j-k-1)\mu_A\big)\\
&\quad+\sum_{j=\tilde{n}_A+1}^{n_A}\Big[\sum_{k=0}^{\tilde{n}_A}\binom{\tilde{n}_A}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{\tilde{n}_A-k}\big((k+1)\mu_D+(\tilde{n}_A-k)\mu_A\big)\\
&\qquad+\sum_{k=0}^{j-\tilde{n}_A-1}\binom{j-\tilde{n}_A-1}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{j-\tilde{n}_A-k-1}\big((k+1)\frac{\mu_D}{2}\\
&\qquad\quad+(j-\tilde{n}_A-k-1)\frac{\mu_A}{2}\big)\Big]\Big]\\
&+(1-\tilde{\gamma}_A)\Big[\sum_{j=1}^{\tilde{n}_D}\sum_{k=0}^{j-1}\binom{j-1}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{j-k-1}\big(k\mu_A+(j-k)\mu_D\big)\\
&\quad+\sum_{j=\tilde{n}_D+1}^{n_D}\Big[\sum_{k=0}^{\tilde{n}_D}\binom{\tilde{n}_D}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{\tilde{n}_D-k}\big(k\mu_A+(\tilde{n}_D-k)\mu_D\big)\\
&\qquad+\sum_{k=0}^{j-\tilde{n}_D-1}\binom{j-\tilde{n}_D-1}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{j-\tilde{n}_D-k-1}\big(k\frac{\mu_A}{2}\\
&\qquad\quad+(j-\tilde{n}_D-k)\frac{\mu_D}{2}\big)\Big]\Big]\Big].
\end{aligned} \tag{2.27}$$

Next we need to compute same metrics but under $\pi = PA$ and $\pi = PD$:

$$E\big[TTFT_A^{PA}|N_A=n_A\big] = \frac{1}{(1-\tilde{\gamma}_D)n_A+\tilde{\gamma}_A(n-n_A)}\times$$

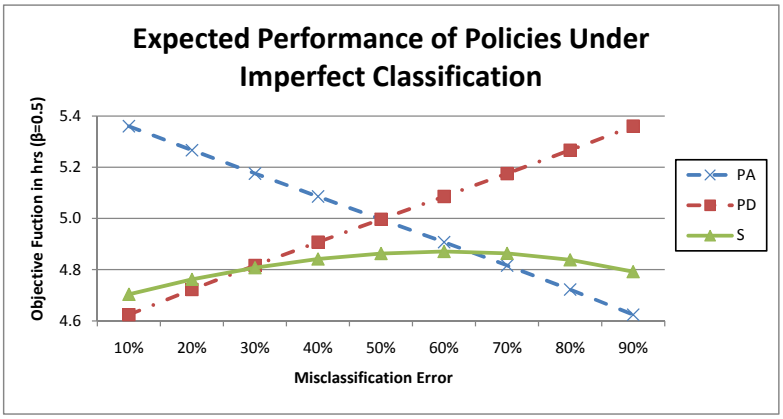


Figure 2.16: Expected performance of policies for a clearing system with $n = 20$, $\mu_A = 80(mins)$, $\mu_D = 45(mins)$, and symmetric misclassification error between A and D patients. Streaming is more robust to misclassification errors than pooling.

$$\times\Big[(1-\tilde{\gamma}_D)\sum_{j=1}^{n_A}\sum_{k=0}^{j-1}\binom{j-1}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{j-k-1}\big(k\frac{\mu_D}{2}+(j-k-1)\frac{\mu_A}{2}\big)$$
$$+\tilde{\gamma}_A\sum_{j=1}^{n-n_A}\Big[\sum_{k=0}^{j-1}\binom{j-1}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{j-k-1}\big(k\frac{\mu_A}{2}+(j-k-1)\frac{\mu_D}{2}\big)$$
$$+\sum_{k=0}^{n_A}\binom{n_A}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{n_A-k}\big(k\frac{\mu_D}{2}+(n_A-k)\frac{\mu_A}{2}\big)\Big]\Big].$$

Moreover, we have:

$$\overline{TTFT}_A^{PA}=\sum_{n_A=0}^{n}E\big[TTFT_A^{PA}|N_A=n_A\big]\times Pr(N_A=n_A),$$

where $Pr(N_A = n_A)$ is given in (2.22).

Similarly we can compute $\overline{LOS}_D^{PA}$:

$$E\big[\overline{LOS}_D^{PA}|N_A=n_A\big]=\frac{1}{\tilde{\gamma}_D n_A+(1-\tilde{\gamma}_A)(n-n_A)}\times$$

$$\Big[\tilde{\gamma}_D \sum_{j=1}^{n_A}\sum_{k=0}^{j-1}\binom{j-1}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{j-k-1}\big((k+1)\frac{\mu_D}{2}+(j-k-1)\frac{\mu_A}{2}\big)$$
$$+(1-\tilde{\gamma}_A)\sum_{j=1}^{n-n_A}\Big[\sum_{k=0}^{j-1}\binom{j-1}{k}\tilde{\gamma}_A^k(1-\tilde{\gamma}_A)^{j-k-1}\big(k\frac{\mu_A}{2}+(j-k)\frac{\mu_D}{2}\big)$$
$$+\sum_{k=0}^{n_A}\binom{n_A}{k}\tilde{\gamma}_D^k(1-\tilde{\gamma}_D)^{n_A-k}\big(k\frac{\mu_D}{2}+(n_A-k)\frac{\mu_A}{2}\big)\Big]\Big], \tag{2.28}$$

and:

$$\overline{LOS}_D^{PA}=\sum_{n_A=0}^{n}E\big[LOS_D^{PA}|N_A=n_A\big]\times Pr(N_A=n_A).$$

It remains to compute the metrics under $\pi=PD$:

$$E\big[TTFT_A^{PD}|N_A=n_A\big]=\frac{1}{(1-\tilde{\gamma}_D)n_A+\tilde{\gamma}_A(n-n_A)}\times$$

$$\Big[\tilde{\gamma}_A\sum_{j=1}^{n-n_A}\sum_{k=0}^{j-1}\binom{j-1}{k}(1-\tilde{\gamma}_A)^k\tilde{\gamma}_A^{j-k-1}\big(k\frac{\mu_D}{2}+(j-k-1)\frac{\mu_A}{2}\big)$$
$$+(1-\tilde{\gamma}_D)\sum_{j=1}^{n_A}\Big[\sum_{k=0}^{j-1}\binom{j-1}{k}(1-\gamma_D)^k\gamma_D^{j-k-1}\big(k\frac{\mu_A}{2}+(j-k-1)\frac{\mu_D}{2}\big)$$
$$+\sum_{k=0}^{n-n_A}\binom{n-n_A}{k}(1-\tilde{\gamma}_A)^k\tilde{\gamma}_A^{n-n_A-k}\big(k\frac{\mu_D}{2}+(n-n_A-k)\frac{\mu_A}{2}\big)\Big]\Big],$$

and:

$$\overline{TTFT}_A^{PD}=\sum_{n_A=0}^{n}E\big[TTFT_A^{PD}|N_A=n_A\big]\times Pr(N_A=n_A).$$

Similarly, we have:

$$E\big[LOS_D^{PD}|N_A=n_A\big]=\frac{1}{\tilde{\gamma}_D n_A+(1-\tilde{\gamma}_A)(n-n_A)}\times$$

$$\Big[(1-\tilde{\gamma}_A)\sum_{j=1}^{n-n_A}\sum_{k=0}^{j-1}\binom{j-1}{k}(1-\tilde{\gamma}_A)^k\tilde{\gamma}_A^{j-k-1}\big((k+1)\frac{\mu_D}{2}+(j-k-1)\frac{\mu_A}{2}\big)$$
$$+\tilde{\gamma}_D\sum_{j=1}^{n_A}\Big[\sum_{k=0}^{j-1}\binom{j-1}{k}\tilde{(}1-\tilde{\gamma}_D)^k\gamma_D^{j-k-1}\big(k\frac{\mu_A}{2}+(j-k)\frac{\mu_D}{2}\big)$$
$$+\sum_{k=0}^{n-n_A}\binom{n-n_A}{k}(1-\tilde{\gamma}_A)^k\tilde{\gamma}_A^{n-n_A-k}\big(k\frac{\mu_D}{2}+(n-n_A-k)\frac{\mu_A}{2}\big)\Big]\Big],$$

and:

$$\overline{LOS}_D^{PD}=\sum_{n_A=0}^{n}E\big[LOS_D^{PD}|N_A=n_A\big]\times Pr(N_A=n_A).$$

Therefore, we have computed expected values of all metrics under different possible policies. Using these computation, Figure 2.16 depicts the performances for a typical numerical example. An important observation is that streaming is much more robust to misclassification errors than the pooling policies.

2.10 Appendix C: Further Descriptions of the Simulation Framework and Assumptions.

In this section we describe the patient flow and assumptions of our simulation framework in more details. Many assumptions are made to be as close as possible to the practice observed in University of Michigan Emergency Department (UMED). A year of data from UMED is gathered to calibrate the simulation. The simulation was developed in a C++ framework. Our model can be described as a cycle-stationary model with a period of one week. Each data point is obtained for 5000 replications of simulating a week, where each replication is preceded by a warm up period of one week (which was observed to be a sufficient warm up period because correlations in the ED flow are small for spans of two or more days). The number of replications (5000) is chosen so that the confidence intervals are tight enough that (1) the sample

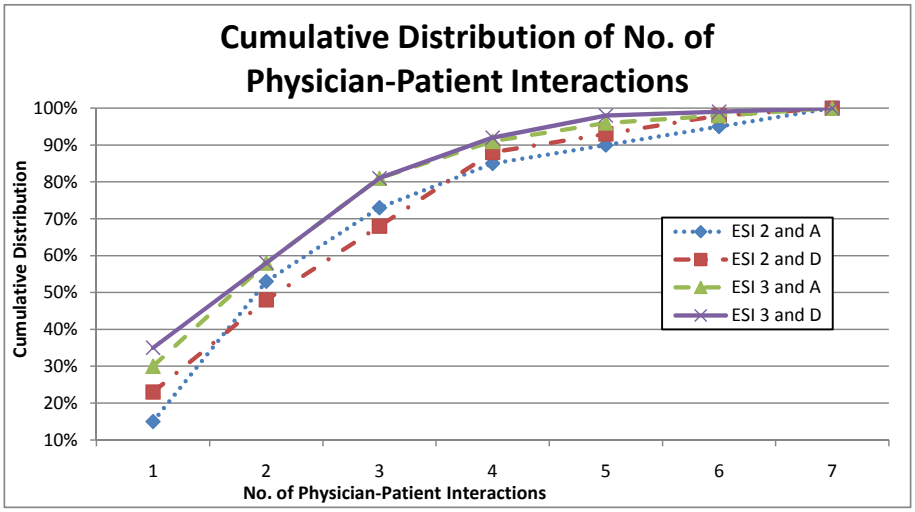


Figure 2.17: Cumulative number of class based physician-patient interactions

averages are reliable, and (2) our data presentation need not to visualize these very tight intervals.

Arrival Process. Arrivals for patient classes are modeled using non-stationary Poisson processes. The arrival rates for different classes (obtained from a year of UMED data) are depicted in Figure 3.5. The general pattern is similar to those found in other studies (e.g., Green et al. (2006)). A "thinning" mechanism (see Lewis and Shedler (1979a) and Lewis and Shedler (1979b)) is used to simulate the non-stationary Poisson process arrivals for each class of patients (with rates depicted in Figure 3.5).

Service Process. The service process in the ED is depicted in Figure 2.5. Each patient goes through several phases of patient-physician interactions/treatment followed by tests and preparations. The duration of each interaction is stochastic and depends on the class of the patient and the number of previous interactions. For instance, the first and last interactions are usually longer than intermediate ones. Also, the duration of "wait" states is stochastic and depend on the class of the patient, based on the information at the UMED. For instance, the last "wait" state, i.e., where the patient is given final directions and is waiting to be disposed is much longer for

admits since they have to be boarded until a bed becomes available in the hospital (the so-called hospital bed-block effect). The number of interactions with a physician per patient ranges from 1 to 7 and depends on the class of the patient, as well as several other factors. Based on the class of the patient, we draw the number of such interactions from a distribution constructed from a detailed time study published in [52] (see Table 3 there) after modifying the data to represent our four patient classes. These class based distributions are depicted in Figure 3.6. The simulated service process is non-collaborative (an ED physician rarely transfers his/her patients to another physician) and non-preemptive (an ED physician rarely moves to another patient in the middle of his/her current interaction).

Phase 1: Assigning Patients to Rooms and Physicians. Whenever a room/bed becomes available, the nurse who is in charge of bed assignment transfers a triaged patient from the waiting area to that room. S/he uses a Phase 1 sequencing rule to decide which patient to bring in to an exam room from the main waiting area (see the body of the paper for different Phase 1 rules implemented). In the VS designs, if an A(D) bed becomes available, the nurse in charge brings an A(D) patient (with priority to patients of ESI 2) from the waiting area in to one of the rooms. If however, an A(D) patient is not waiting in the waiting area, the nurse brings in a D(A) patient (with priority to patients of ESI 2). Also, after an A(D) patient is triaged, s/he is directly guided to one of the A(D) beds if one such bed is available, and if not, to one of the D(A) beds (i.e., bed sharing is allowed, since beds are only virtually separated). If, however, no bed is available, the patient has to wait in the waiting area. Once a room/bed is assigned to a patient, the bed cannot be occupied by another patient until s/he leaves the ED; the bed assigned to a patient cannot be assigned to any other one, even if the patient is sent to another facility for a test. After the patient is brought into the room, s/he goes through the first "waiting" state (i.e., initial preparation by a nurses) which takes some stochastic amount of time. The average duration of this stage depends on the class of the patient. After

this stage the patient is assigned to a physician (if a physician is available) where his/her first treatment starts. The rule to choose a physician is generally to assign the patient to the physician who is handling the lowest number of patients (among those available at that time). However, the rules to choose a physician is different between the virtual streaming (VS) and the pooling patient flow designs, since in a VS design the physicians are divided to two groups one for A patients and one for D patients. Under a VS design, if the patient is assessed to be of A(D) type, the priority is given to physicians devoted to A(D). In other words, an available A(D) type physician is allowed to cross to the other stream only if a physician of D(A) type is needed but is not available (due to being busy with a patient or being currently assigned to the maximum number of patients that a physician is willing to handle). Under pooling designs, physicians do not have labels and therefore a physician who is handling the lowest number of patients (among those available at that time) becomes responsible for the newly arrived patient. Once a physician is assigned to a patient s/he is the only physician who can work on that patient. If no physician is available at the time the patient is ready for his/her first interaction with the physician, the patient has to wait in the exam room.

Phase 2: Which patient to choose next? Whenever a physician finishes a treatment stage (including direct and indirect interactions), s/he is available to visit another patient. The physician chooses the next patient based on the instructions s/he is given according to the Phase 2 sequencing rule. If the physician has less than the upper bound on the number of patients that a physician is willing to handle (7 was used based on the UMED data), s/he can also choose to initialize a new journey by taking a new patient: visiting a patient who has been taken to a room but has been waiting for a physician to become available. Under the VS designs, physicians with A(D) label first use the Phase 2 priority rule on the patients of A(D) type and are allowed to handle D(A) patients only to avoid starvation.

References (for the Appendixes of this chapter)

Green, L. V., J. Soares, J. F. Giglio, R.A. Green. 2006. Using queuing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* 13(1) 61-68.

Lewis, P. A.W., G. S. Shedler. 1979a. Simulation of nonhomogenous poisson processes by thinning. *Naval Research Logistics Quarterly* 26(3) 403-413.

Lewis, P. A.W., G. S. Shedler. 1979b. Simulation of nonhomogenous poisson processes with degree-two exponential polynomial rate function. *Oper. Res.* 27(5) 1026-1039.

CHAPTER 3

Emergency Department: Complexity-Based Triage

3.1 Introduction

Overcrowding and lapses in patient safety are prevalent problems in Emergency Departments (ED's) in the U.S. and around the world. In one study, 91% of U.S. ED's responding to a national survey reported that overcrowding was a problem, and almost 40% of them reported overcrowding as a daily occurrence ([5]). In addition to causing long wait times, many research studies have linked delays due to overcrowding to elevated risks of errors and adverse events (see, e.g., [143], [51], [148], and [100]). This situation prompted the Institute of Medicine's Committee on Future of Emergency Care in the United States Health System to recommend that *"hospital chief executive officers adopt enterprisewide operations management and related strategies to improve the quality and efficiency of emergency care"* ([77]). The triage process is a natural place to introduce operations management (OM) into the ED.

Triage (a word derived from the French verb "trier," meaning "to sort") refers to the process of sorting and prioritizing patients for care. [47] argue that there are two main purposes for triage: *"[1] to ensure that the patient receives the level and quality of care appropriate to clinical need (clinical justice) and [2] that departmental resources are most usefully applied (efficiency) to this end."* (see [109] for further

discussion of the underlying principles and goals of triage).

While current triage systems used around the world address the clinical justice purpose of triage, the efficiency purpose has been largely overlooked. For instance, most ED's in Australia use the Australasian Triage Scale (ATS), the Manchester Triage Scale (MTS) is prevalent in the U.K., and ED's in Canada generally use the Canadian Triage Acuity Scale (CTAS). While they differ in their details, all of these triage systems classify patients strictly in terms of urgency and so address only the first (clinical justice) purpose of triage.

In the U.S., many ED's continue to use a traditional urgency-based 3-level triage scale, which categorizes patients into emergent, urgent, and non-urgent classes. But other U.S. hospitals have adopted the 5-level Emergency Severity Index (ESI) system (see [45]), which combines urgency with an estimate of resources (e.g., tests) required. In the ESI system (a typical version of which is illustrated in Figure 3.1 (left)), urgent patients who cannot wait are classified as ESI-1 and 2, while non-urgent patients who can wait are classified as ESI-3, 4, and 5. ESI-4 and 5 patients are usually directed to a fast track (FT) area, while ESI-1 patients are immediately moved to a resuscitation unit (RU). ESI-2 and 3 patients, who represent the majority of patients at large academic hospitals (e.g., about 80% at the University of Michigan ED (UMED)), are served in the main area of the ED with priority given to ESI-2 patients. Since the ESI system does not differentiate between patients in the ESI-2 and ESI-3 categories in terms of complexity, patients in the main ED are still sorted and prioritized purely on the basis of urgency. Hence, the ESI system does not respond to the second purpose of triage for the majority of the patients. As [160] state, *"Many clinicians have already realized that triage as it is widely practiced today no longer meets the requirement of timely patient care."* Our goal in this chapter is to propose a new triage system, which we call *complexity-based triage*, that can significantly improve ED performance with respect to both clinical justice and efficiency.

Doing this poses two challenges: (a) deciding what information should be collected

at the time of triage, and (b) determining how this information should be used to assign patients to tracks and prioritize them within tracks (see, e.g., [90]). [121] proposed that one way ED's can improve performance is to have triage nurses predict the final disposition (admit or discharge) of patients in addition to assigning an ESI level. Assigning patients to separate admit and discharge streams can reduce average time to first treatment for admit patients and average length of stay for discharge patients. But this study also indicated that the performance of the streaming policy improves as the difference between the average treatment times of admit and discharge patients becomes larger. This suggests that classifying patients according to complexity may be even more useful than classifying them according to ultimate disposition.

There is ample evidence from the OM literature that classifying patients based on their service requirements and giving priority to those with shorter service times (e.g., by following a Shortest Processing Time (SPT) priority rule) can improve resource usage efficiency, and thereby reduce the average waiting time among all patients. Furthermore, empirical studies from the emergency medicine literature suggest that patients can be effectively classified by complexity at the time of triage. Specifically, [153] defined complex patients as those requiring at least two procedures, investigations, or consultations and concluded that *"Triage nurses are able to make valid and reliable estimates of patient complexity. This information might be used to guide ED work flow and ED casemix system analysis."*

Using the number of (treatment related) interactions with the physician (which correlates directly with expected treatment duration) as an indicator of patient complexity, we propose and investigate the benefit of the new complexity-based triage process depicted in Figure 3.1 (right). Note that, unlike the ESI system, our proposed system classifies all patients (except those at risk of death) in terms of complexity. In this chapter, we compare our proposed triage system with current urgency-based systems and show that incorporating patient complexity into the triage process can yield substantial performance benefits. To do this, we consider ED performance in

terms of both risk of adverse events (clinical justice) and average length of stay (efficiency). Specifically, we make use of a combination of analytic and simulation models calibrated with hospital data to examine the following:

1. **Prioritization:** *How should ED's use complexity-based triage information to prioritize patients?*
2. **Magnitude:** *How much benefit does complexity-based triage (which adds complexity information to conventional urgency evaluations) offer relative to urgency-based triage?*
3. **Sensitivity:** *How sensitive are the benefits of complexity-based triage to misclassification errors and other characteristics that may vary across ED's?*
4. **Design:** *Should complexity-based information be used to create separate service streams for simple and complex patients, or is it better to use it to prioritize patients in a traditional pooled flow design?*

In addition to collecting detailed ED data (from UMED), addressing these practical questions required us to make some technical innovations: (1) In the ED, upfront triage misclassifications are inevitable. However, the literature on priority queueing systems under misclassification is very limited. We contribute to this literature by explicitly considering misclassifications and deriving optimal control policies under different settings that effectively approximate the ED environment. We do this through a linear transformation of control indices so that they represent "error-impacted" rates, which use only information from historical data. This leads to modified versions of the well-known $c\mu$ rule, which we show to be very effective as the basis for prioritizing patients into ED examination rooms. (2) To provide guidance for ED physicians on how to prioritize patients within the examination rooms (when they have a choice of what patient to see next), we develop a Markov Decision Process (MDP) model. A challenging feature of this model, which is common in many other

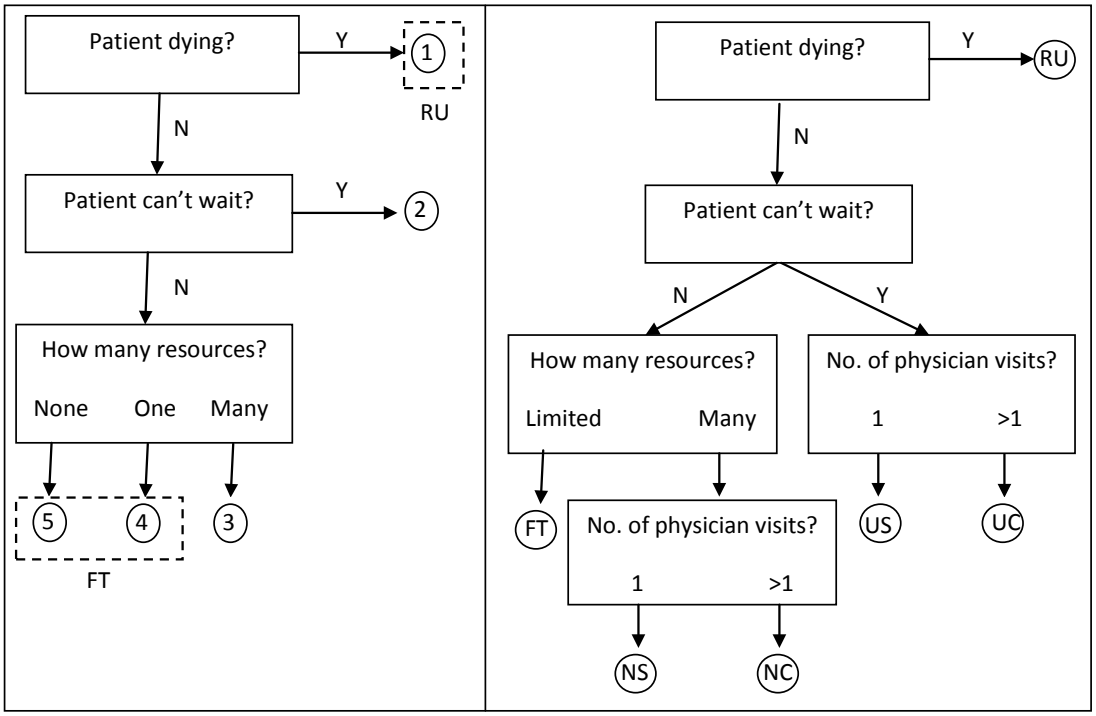


Figure 3.1: Left: Current practice of triage (Emergency Severity Index (ESI) algorithm version 4); Right: Proposed complexity-based triage system (RU: Resuscitation Unit, FT: Fast Track, NS: Non-urgent Simple, NC: Non-urgent Complex, US: Urgent Simple, UC: Urgent Complex).

heath delivery settings, is that patients are occasionally sent for tests (e.g., MRI, CT Scan, X-Ray, etc.), and are unavailable to the physician during testing. In such a setting, the physician (controller) may need to consider both the current and the future availability of the patients when making decisions. This type of problems usually result in complex state-dependent optimal control policies. However, we show how a simple-to-implement rule that relies only on historical data defines the optimal policy for ED physicians. (3) Because of unbounded transition rates, the MDP model of patient prioritization within examination rooms cannot use the conventional method of uniformization (proposed by [99]) for working with continuous-time MDP's. The available technical results for continuous-time MDP's with unbounded transition rates is very limited (see, e.g., [56]). We contribute to this literature by showing how one can use a sequence of MDP's, each with bounded transition rates,

to derive an optimal policy for the original MDP. Using this innovative technique, we derive a simple-to-implement rule for ED physicians that prescribes which patient to visit next.

The remainder of the chapter is organized as follows. Section 3.2 summarizes previous OM and medical research relevant to our research questions. Section 6.3 describes our performance metrics and analytical modeling approach. For modeling purposes, we divide the ED experience of the patient into Phase 1 (from arrival until assignment to an examination room) and Phase 2 (from assignment to an examination room until discharge/admission to the hospital). Section 3.4 focuses on Phase 1 and uses analytical queueing models to compare performance under urgency-based and complexity-based triage systems. Section 3.5 considers Phase 2 by developing and analyzing a Markov Decision Process model. Section 3.6 uses a high-fidelity simulation model of the full ED to validate the insights obtained through our analytical models and to refine our estimates of the magnitude of performance improvement possible with complexity-based triage. We conclude in Section 4.5.

3.2 Literature Review

In this section, we review studies related to our work from both the operations research/management literature and the medical literature.

3.2.1 Operations Research/Management Studies

The effect of assigning priorities in queueing systems has been studied in the operations research literature for a long time. One of the first works to rigorously analyze such systems under perfect classification was [28]. Assuming perfect customer classification, [28] and [29] showed that the expected waiting time among all customers can be reduced by assigning priorities. [149] extended Cobham's results to the case with imperfect classification for a two-priority single-channel system. They

recommended creating a "mixed" group for customers who cannot be classified with certainty into either group 1 or 2, and assigning priorities probabilistically within this group. Further analysis of priority queueing systems can be found in [31], [83], and [162].

Under perfect classification, average holding cost objective, Poisson arrivals, and a non-preemptive non-idling single server model, [31] used an interchange argument to show that the $c\mu$ rule is optimal among priority rules. That is, product of the holding cost rate times the service rate is the index that quantifies the attractiveness/priority of that job or job class. [86] extended this result and used a semi-Markov decision process to show that the $c\mu$ rule remains optimal even among the larger class of state-dependent policies with or without the option of idling the server. The $c\mu$ rule has since been shown to be optimal in many other queueing frameworks; see, e.g., [22], [151], [155], [124], and references therein. In this chapter, we contribute to this literature by proving the optimality of modified versions of the $c\mu$ rule that use "error-impacted" indices, which are well-suited to the ED triage environment where misclassification is inevitable.

Research related to our work that analyzes the performance of ED's from an operations perspective is also very limited. [121] considered streaming of ED patients based on triage estimations of the final disposition (admit or discharge) and found that an appropriate "virtual streaming" policy can improve performance with respect to the operational characteristics of average Length of Stay (LOS) and Time To First Treatment (TTFT). [130] considered the impact of non-emergency patients on ED delays using urgency-based triage, and proposed a simple priority queueing model to reduce average waiting times. [157] considered a queue of heterogenous high risk patients, for which treatment times are exponential, and patient classification is perfect, and concluded that patients should be prioritized into as many urgency classes as possible in order to maximize survival. [11] used the average waiting time as the performance metric in a service system with two classes of customers, in which cus-

tomer classification is imperfect, and showed that prioritizing customers according to the probability of being from the class that should have a higher priority when classification is perfect outperforms any finite-class priority policy.

The above studies suggest that separating patients according to a measure of service duration can reduce waiting times through a better resource allocation. However, we note that they (a) lack insights into clinical justice/safety issues that are vital in ED's, and (b) are limited to simple/stylized queueing models with features (e.g., one-stage service, fixed number of customers all available at time zero, availability of the customers at any time during service, no bound on the number of customers that can be assigned to a server, no change in the condition/holding cost of customers after they begin service, perfect customer classification, etc.) that do not capture the reality of ED's. In this chapter, we seek to address both safety and efficiency, and to account for the key features that define the ED environment. To this end, in addition to using stylized models that approximate ED flow, we develop a complex simulation model of the ED and use hospital data to investigate whether the insights from stylized models carry over to the actual ED environment.

3.2.2 Medical Studies

Our research was informed by empirical studies of ED's and triage processes. [50], [47], and [76] provide excellent reviews of the history of the triage process and its development over time. Most studies attribute the first formal battlefield triage system to the distinguished French military surgeon Baron Dominique-Jean Larrey who recognized a need to evaluate and categorize wounded soldiers. He recommended treating and evacuating those requiring the most urgent medical attention, rather than waiting hours or days for the battle to end before treating patients, as had been done in previous wars ([76]). Since that time, triage in medicine has been mainly based on urgency. However, the idea of considering the complexity of patients goes back to World War I triage recommendations: *"A single case, even if it urgently*

requires attention, –if this will absorb a long time,– may have to wait, for in that same time a dozen others, almost equally exigent, but requiring less time, might be cared for. The greatest good of the greatest number must be the rule." ([87]). The ESI triage system shown in Figure 3.1 (left) is the most serious effort to date at introducing complexity into the triage process. However, because (a) the number of resources required does not necessarily correlate with the physician time required by the patient, (b) The complexity of patients varies greatly within ESI categories, and (c) ED's do not use ESI information in a consistent manner to prioritize patients, the ESI system falls well short of the potential for complexity-based triage.

Anticipating the potential of complexity-based triage, [153] empirically tested the ability of nurses to estimate patient complexity at the time of triage and found that they are able to this reliably. [153] suggested that this type of information could be used to improve patient flow in ED's, although they did not specify how. Other researchers have suggested using physicians at triage as a way to generate more and better patient information. However, [60] and [119] studied physician triage and found that it is not an effective method for reducing total length of stay, although it may reduce the average time spent in an ED bed.

Finally, several studies have been published in medical journals that aim at investigating and/or validating the ESI triage system. For this stream of research, we refer interested readers to [45], which summarized the findings and recommendations of a task force from the American College of Emergency Physicians (ACEP) and the Emergency Nurses Association (ENA) appointed in 2003 to analyze the literature and make recommendations regarding use of 5-level triage systems in the United States. While this committee found the 5-level ESI system to be a good option compared to other available methods, they encouraged further in-depth research for improving the triage system.

3.3 Modeling the ED

To answer the four questions (prioritization, magnitude, sensitivity, and design) we posed in Section 1, we need to model patient flow through the ED. A high level schematic of this flow is presented in Figure 3.2. A patient's path through the ED begins with *arrival*, which occurs in a non-stationary stochastic manner. Upon arrival, the patient goes to *triage*, where s/he is classified according to a predefined process (based on urgency and/or complexity), which inevitably involves some misclassification errors. If an examination room is not immediately available, s/he goes to the ED *waiting* area until s/he is called by the charge nurse and brought to an examination room. There s/he goes through a stochastic number of *treatment* stages with a physician, which are also stochastic in duration. These treatment stages are punctuated by *test* stages which involve testing (MRI, CT Scan, X-Ray etc.) or preparation/processing activities that do not involve the physician and during which the patient is unavailable to the physician. The final processing stage after the last physician interaction is *disposition*, in which the patient is either *discharged* to go home or *admitted* to the hospital.

We refer to the time a patient spends after s/he is triaged and before s/he is brought an examination room as "Phase 1," and label the remainder of time in the ED until disposition as "Phase 2." Because they are under observation and care, patients have a lower risk of adverse events during Phase 2 than during Phase 1. Patients are taken from Phase 1 to Phase 2 by the charge nurse based on a Phase 1 sequencing rule that can make use of the patient classification performed at triage. Similarly, in Phase 2, physicians use some kind of a sequencing rule to choose which patient to see next.

To gain insights into appropriate triage and priority rules, we first focus on the risk of adverse events and average waiting times in Phase 1 by considering the dashed area in Figure 3.2 (i.e., Phase 2) as a single-stage service node with a single, aggregated server. Since ED's rarely send a patient back to the waiting area of Phase 1 once

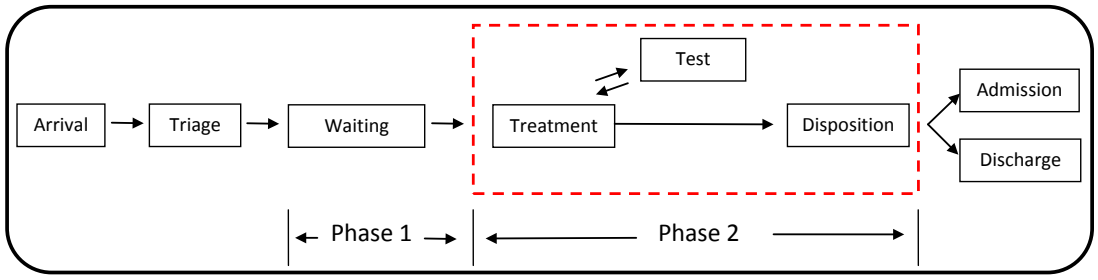


Figure 3.2: General flow of patients in the main ED.

s/he has begun service, we assume a non-preemptive service protocol. We also approximate the non-stationary arrival process by a stationary Poisson process. These simplifications allow us to gain insights into suitable Phase 1 priority rules using a multi-class non-preemptive priority M/G/1 queueing model. We refer to this model as the *simplified single-stage ED model*. An important and challenging aspect of this model is the existence of triage misclassifications that can affect the way patients should be prioritized.

After analyzing this model, we focus on the risk of adverse events and average waiting times in Phase 2. To do this, we note that physicians can preempt their current interaction with a patient to visit another patient with a higher priority (e.g., a severely acute patient), and hence, we allow for preemption in Phase 2. Again approximating arrivals with a stationary Poisson process arrival stream, we can represent the multi-stage service process in Phase 2 as a Markov decision process model, which we label as the *simplified multi-stage ED model*. We use this model to get insights into appropriate Phase 2 priority rules that physicians can implement when choosing their next patient.

Finally, we test the insights from both analytic models under realistic conditions with a high fidelity simulation model of the full ED calibrated with a year of data from University of Michigan Hospital ED as well as time study data from the literature.

3.4 Phase 1: A Simplified Single-Stage ED Model

To formalize the Phase 1 sequencing problem, we define a patient to be of type ij if his/her urgency level is $i \in U$ and his/her complexity type is $j \in C$, where $U = \{U(\text{Urgent}), N(\text{Non-urgent})\}$ and $C = \{C(\text{Complex}), \text{S}(\text{Simple})\}$. We suppose patients of type $ij \in U \times C$ arrive according to a Poison process with rate λ_{ij} and have service times (i.e., the total time spent in Phase 2) that follow a distribution, $F_{ij}(s)$ with first moment $1/\mu_{ij}$ (where $\mu_{iC} \leq \mu_{iS}$ for all $i \in U$) and a finite second moment. We assume patients of type ij are subject to adverse events which occur according to a Poisson process with rate θ_{ij}, where $\theta_{Uj} \geq \theta_{Nj}$ for all $j \in C$. Notice that adverse events only rarely result in death, i.e., the average reported number of adverse events per patient is much higher than the average number of death per patient (see, e.g., [100] where the authors report that 28% of patients boarded in the ED had some adverse event or error in the course of boarding only). Thus, we assume that the service process continues, so that it is possible for a patient to experience more than one adverse event. This allows us to compare the performance of the ED under different triage systems in a systematic way. Similarly, changes in patient priority after the occurrence of an adverse event can be neglected, since (a) such changes are rare, and (b) the effect of such rare changes are not systematically different under different triage systems.

Assuming $R_\pi^\Omega(t)$ represents the counting process that, under patient classification (i.e., triage) policy Ω and sequencing rule π, counts the total number of adverse events (for all patients) until time t, we consider $R_\pi^\Omega = \lim_{t\to\infty} R_\pi^\Omega(t)/t$ (when the limit exists) as our metric and refer to it as the *rate of adverse events (ROAE)*. However, if $\theta_{ij} = 1$ for all $i \in U$ and $j \in C$, then it can be shown that $R_\pi^\Omega / \sum_{i\in U}\sum_{j\in C} \lambda_{ij}$ represents the average *length of stay (LOS)*. (Notice that the sample path costs of LOS and adverse events with unit risk rates divided by total arrival rate will be different, but they are equal in expectation.) Hence, we can use our metric to characterize performance with respect to both safety and efficiency in a systematic and coherent way.

3.4.1 Urgency-Based Triage - Phase 1

We first consider current practice in most ED's, in which patients are classified solely based on urgency, and use our simplified single-stage model to focus on Phase 1 sequencing decisions. We start with the case of perfect classification and then consider the case of stochastic misclassification.

When patients can be perfectly classified as either urgent (U) or non-urgent (N), the arrival rates for U's and N's are $\lambda_U = \sum_{j\in C} \lambda_{Uj}$ and $\lambda_N = \sum_{j\in C} \lambda_{Nj}$, respectively. Similarly, the average service times for U's and N's are $1/\mu_U = \sum_{j\in C}(\lambda_{Uj}/\lambda_U)(1/\mu_{Uj})$ and $1/\mu_N = \sum_{j\in C}(\lambda_{Nj}/\lambda_N)(1/\mu_{Nj})$, respectively. Furthermore, from known results for non-preemptive priority queues (see, [28], [149], Section 3.3 of [31], or Section 10.2 of [162]), the average waiting (queue) time of the kth priority class is

$$W_k = \frac{\lambda\,\mathbb{E}(s^2)}{2(1-\sum_{l<k}\rho_l)(1-\sum_{l\le k}\rho_l)}, \tag{3.1}$$

where $\rho_l = \lambda_l/\mu_l$ for class l. Hence, if U's are prioritized over N's, then the average waiting time is $W_U = \lambda\,\mathbb{E}(s^2)/2(1-\rho_U)$ for U's and $W_N = \lambda\,\mathbb{E}(s^2)/2(1-\rho_U)(1-\rho)$ for N's. Furthermore, the average rate of adverse events for U's is $\theta_U = (\lambda_{US}/\lambda_U)\theta_{US} + (\lambda_{UC}/\lambda_U)\theta_{UC}$ and for N's is $\theta_N = (\lambda_{NS}/\lambda_N)\theta_{NS} + (\lambda_{NC}/\lambda_N)\theta_{NC}$. With these, the ROAE under an urgency-based triage policy (i.e., patient classification with respect to set U) that gives priority to U's is

$$R_U^U = \theta_U\lambda_U(\lambda\mathbb{E}(s^2)/2(1-\rho_U)) + \theta_N\lambda_N(\lambda\mathbb{E}(s^2)/2(1-\rho_U)(1-\rho)). \tag{3.2}$$

Similarly, we can obtain the ROAE under an urgency-based triage policy that gives priority to N's:

$$R_N^U = \theta_N\lambda_N(\lambda\mathbb{E}(s^2)/2(1-\rho_N)) + \theta_U\lambda_U(\lambda\mathbb{E}(s^2)/2(1-\rho_N)(1-\rho)). \tag{3.3}$$

Comparing these reveals that, without misclassification errors, the best priority rule

is to prioritize U's (N's) if, and only if, $\theta_U \mu_U \geq (\leq) \theta_N \mu_N$. Given the criteria used to classify a patient as urgent, we expect θ_U and θ_N be such that $\theta_U \mu_U > \theta_N \mu_N$, meaning that U's will be given priority. However, this simple result may or may not hold if one carefully considers the effect of stochastic triage misclassifications.

Therefore, we now formally incorporate stochastic misclassification errors into our decision model of urgency-based triage and prioritization. Let γ_U and γ_N denote the misclassification probabilities for urgent and non-urgent patients, respectively. The arrival rates for patients classified (correctly or erroneously) as U and N are $\lambda'_U = \lambda_U(1-\gamma_U) + \lambda_N \gamma_N$ and $\lambda'_N = \lambda_N(1-\gamma_N) + \lambda_U \gamma_U$, respectively. Similarly, the mean service times for patients classified as U and N are $1/\mu'_U = [\lambda_U(1-\gamma_U)(1/\mu_u) + \lambda_N \gamma_N (1/\mu_N)]/\lambda'_U$ and $1/\mu'_N = [\lambda_N(1-\gamma_N)(1/\mu_N) + \lambda_U \gamma_U (1/\mu_U)]/\lambda'_N$, respectively. Finally, the ROAE for patients classified as U and N are $\theta'_U = [\lambda_U(1-\gamma_U)\,\theta_U + \lambda_N\,\gamma_N\,\theta_N]/\lambda'_U$ and $\theta'_N = [\lambda_N(1-\gamma_N)\,\theta_N + \lambda_U \gamma_U\,\theta_U]/\lambda'_N$, respectively.

Using (3.2) with these new "error impacted" rates shows that when priority is given to U's, the ROAE under imperfect classification is:

$$R_U^{U'} = \theta'_U \lambda'_U (\lambda \mathbb{E}(s^2)/2(1-\rho'_U)) + \theta'_N \lambda'_N (\lambda \mathbb{E}(s^2)/2(1-\rho'_U)(1-\rho)), \quad (3.4)$$

where $\rho'_U = \lambda'_U/\mu'_U$. Similarly, using (3.3) shows that when priority is given to N's:

$$R_N^{U'} = \theta'_N \lambda'_N (\lambda \mathbb{E}(s^2)/2(1-\rho'_N)) + \theta'_U \lambda'_U (\lambda \mathbb{E}(s^2)/2(1-\rho'_N)(1-\rho)), \quad (3.5)$$

where $\rho'_N = \lambda'_N/\mu'_N$.

The above results enable us to state:

Proposition 4 (Phase 1 Prioritization - Urgency-Based Triage) *In the simplified single-stage ED model with imperfect urgency-based classification:*
(i) The best priority rule is to prioritize U patients if $\theta'_U\,\mu'_U \geq \theta'_N \mu'_N$; otherwise, prioritize N patients.
(ii) The best priority rule is the same as that for the case without misclassification

error if $\gamma_N + \gamma_U \leq 1$; otherwise, the best priority ordering is reversed.

Empirical studies have observed misclassification levels γ_N and γ_U to be in the range 9-15% depending on the level of triage nurse experience ([63]). Thus, if, as we expect, prioritizing urgent patients is optimal when there is no misclassification error, the above proposition implies that doing so remains optimal even under realistic levels of misclassification errors. Hence, prioritizing ESI-2 patients over ESI-3 patients in the main ED seems legitimate in the current urgency-based triage practice in the U.S. However, we note that there is wide variance of complexity among ESI-2 and ESI-3 patients. Hence, if complexity is taken into account, simply prioritizing ESI-2 patients over ESI-3 patients as is currently done in practice for these majority of ED patients may be significantly suboptimal. We investigate this issue in the next section.

3.4.2 Complexity-Based Triage - Phase 1

We now consider the complexity-based triage policy shown in Figure 3.1 (right), and compare its performance with respect to that of urgency-based triage currently in use in practice. By doing this we seek to gain insights into the prioritization, magnitude, and sensitivity questions posed in the Introduction.

To evaluate the performance of complexity-based triage when classification is imperfect, we let γ_U and γ_N denote the misclassification error rates with respect to set U. That is, γ_U and γ_N denote the probabilities of classifying a U patient as an N, and an N patient as a U, respectively. Similarly, let γ_C and γ_S denote the misclassification error rates with respect to set C; γ_C denotes the probability that a C patient is classified as an S, and γ_S denotes the probability that an S patient is classified as a C. We assume the misclassification probabilities with respect to sets U and C are independent. As noted earlier, misclassification error rates in terms of urgency have been observed to be in the range of 9-15% ([63]). [153] have tested the ability of triage nurses to evaluate patient complexity (where complexity is defined as requiring two

or more procedures, investigations, or consultation) and observed a misclassification rate of 17%.

Similar to what we did in Section 3.4.1, we need to calculate the error impacted rates λ'_{ij}, θ'_{ij}, and μ'_{ij}. Let $\underline{\lambda} = (\lambda_{US}, \lambda_{UC}, \lambda_{NS}, \lambda_{NC})$ and $\underline{\lambda'} = (\lambda'_{US}, \lambda'_{UC}, \lambda'_{NS}, \lambda'_{NC})$. Then $\underline{\lambda'}$ can be obtained through a linear transformation of $\underline{\lambda}$; $\underline{\lambda'}^T = A\,\underline{\lambda}^T$, where A is a (known) *misclassification error matrix*, and is defined as

$$A = \begin{pmatrix} (1-\gamma_U)(1-\gamma_S) & (1-\gamma_U)\gamma_C & \gamma_N(1-\gamma_S) & \gamma_N\gamma_C \\ (1-\gamma_U)\gamma_S & (1-\gamma_U)(1-\gamma_C) & \gamma_N\gamma_S & \gamma_N(1-\gamma_C) \\ \gamma_U(1-\gamma_S) & \gamma_U\gamma_C & (1-\gamma_N)(1-\gamma_S) & (1-\gamma_N)\gamma_C \\ \gamma_U\gamma_S & \gamma_U(1-\gamma_C) & (1-\gamma_N)\gamma_S & (1-\gamma_N)(1-\gamma_C) \end{pmatrix}. \tag{3.6}$$

Similarly, if $\underline{\theta'}$ and $\underline{\mu'}$ denote the vector of error impacted adverse event and service rates, we have $\underline{\theta'}^T = (A\,(\underline{\lambda} \times \underline{\theta})^T)/\underline{\lambda'}$ and $(\underline{1}/\underline{\mu'})^T = (A\,(\underline{\lambda}/\underline{\mu})^T)/\underline{\lambda'}$, where $\underline{1} = (1,1,1,1)$ and operators "$\times$" and "/" are componentwise multiplier and division, respectively.

With these, the waiting times for each customer class under an imperfect $U \cup C$ classification can be computed using (3.1) with rates replaced with their transformed error impacted counterparts. This model permits us to show the following.

Proposition 5 (Phase 1 Prioritization - Complexity-Based Triage) *In the simplified single-stage ED model with imperfect urgency and complexity classifications:*
(i) The best priority rule is to prioritize patients in decreasing order of $\theta'_{ij}\,\mu'_{ij}$ values.
(ii) $R_*^{U' \cup C'} \leq R_*^{U'}$. That is, even with misclassification errors, implementing the best priority rule for complexity-based triage is always (weakly) better than the optimal priority rule for urgency-based triage.
(iii) The best priority rule of part (i) is optimal even among the larger class of all non-anticipative (state or history dependent, idling or non-idling, etc.) policies.

Proposition 5 (i) addresses the prioritization question by suggesting a simple prior-

ity rule (analogous to the well-known "$c\mu$" rule) to incorporate complexity information into Phase 1 sequencing. Proposition 5 (ii) begins to address the magnitude question by suggesting that complexity-based triage outperforms urgency-based triage, given that the optimal priority rule is implemented. While priority rules are greedy and usually suboptimal, part (iii) confirms that they are optimal in this setting. The surprise is that it is never optimal to idle when only low priority patients are available, even though the model disallows preemption. Furthermore, part (iii) of Proposition 5 states that a dynamic (i.e., state-dependent) priority policy cannot beat the greedy and simple state-independent policy presented in part (i).

Figure 3.3 provides additional insights into the magnitude question by illustrating the amount of improvement for a numerical example with $\mu_{UC} = \mu_{NC} = \mu_C = 1$, $\mu_{US} = \mu_{NS} = \mu_S$ varying from 2 to 5, $\lambda_{US} = (1/5)\mu_{US}, \lambda_{UC} = 1/4, \lambda_{NS} = (1/3)\mu_{NS}, \lambda_{NC} = 1/6, \mathbb{E}(s^2) = 4, \theta_{NS} = \theta_{NC} = \theta_N = 1, \theta_{US} = \theta_{UC} = \theta_U$. Note that (1) the amount of improvement is depicted both in terms of average length of stay and risk of adverse events (since when $\theta_U/\theta_N = 1$, the percentage improvement in risk of adverse events and length of stay are equal), and (2) reduction in the length of stay results in reduction in congestion (by Little's Law), which can serve as a potent remedy for the prevalently observed phenomenon of ED overcrowding. Figure 3.3 suggests that, if the average service time of complex patients is 3–4 times larger than that of simple patients, then complexity-based triage can reduce the risk of adverse events (ROAE) and average length of stay (LOS) by 12–22% and 27–33%, respectively. Finally, we can address the sensitivity question by using our model to determine the environmental factors that favor complexity-based triage.

Proposition 6 (Attractiveness of Complexity-Based Triage) *Under the simplified single-stage ED model, complexity-based triage is more beneficial in ED's with (i) higher utilization, (ii) higher heterogeneity in the average service time of simple and complex patients, (iii) a more equal fraction of simple and complex patients, and (iv) lower error rates in classifying simple and complex patients.*

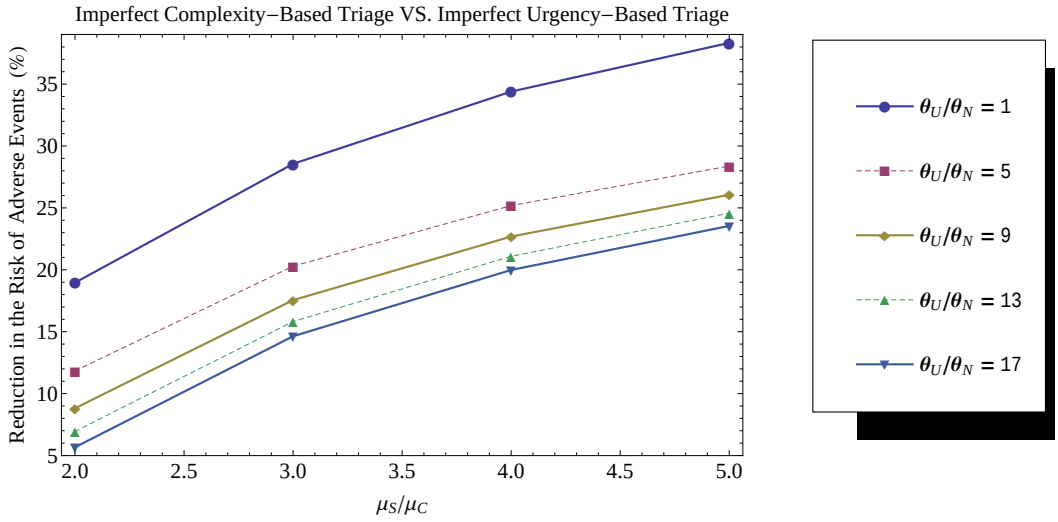


Figure 3.3: Benefit of complexity-based triage over urgency-based triage with practical misclassification rates ($\gamma_{\mathbf{U}} = \gamma_{\mathbf{N}} = \mathbf{10}\%, \gamma_{\mathbf{S}} = \gamma_{\mathbf{C}} = \mathbf{17}\%$) reported in the literature

3.5 Phase 2: A Multi-Stage ED Model

The analysis of the previous section was limited to patient waiting and risk of adverse events prior to entry into an examination room (i.e., Phase 1). But, as illustrated in Figure 3.4, a great deal of ED activity takes place after this point, which contributes to both patient length of stay and risk of adverse events. Since triage classification can be used to sequence patients within the ED, as well as in the waiting room, it is important to consider Phase 2 sequencing as part of our evaluation of complexity-based triage. It seems intuitive that a priority rule similar to that for Phase 1 should serve as a useful guide to physicians in allocating their time among their slate of patients. However, investigating this requires a model of Phase 2 that includes some challenging new features (e.g., patients going off for tests, multiple patients of different class in different stages of treatment, etc.), which did not exist in the model of Phase 1.

To formulate a suitable model, consider the multi-stage service process illustrated

in Figure 3.4 and suppose patients of class $ij \in U \times C$ arrive according to a Poison process with rate λ_{ij}. Further, we suppose the rate of adverse events in Phase 2 is denoted by the vector $\underline{\hat{\theta}} = (\hat{\theta}_{ij})_{ij \in U \times C}$ (which is usually less than the risk of adverse events in Phase 1, $\underline{\theta}$, because patients are monitored and treated in the examination rooms). As they enter examination rooms, patients are assigned to physicians who treat them, often with multiple visits, until their discharge or admission to the hospital. Since an individual physician may be assigned to several patients s/he often has a choice about who to see next among his/her available patients. We call patients who have completed tests (e.g., MRI, CT Scan, X-Ray, etc.) and have results and are ready for a physician visit "available," and patients being tested, prepared, or waiting for results "unavailable." Our model for the congestion for tests is exogenous and can be estimated from historical data. An important feature in modeling the physician choice is the uncertain duration of unavailability of patients to the physician due to the wait for tests and their results.

Suppose each interaction with a patient of class ij takes an exponentially distributed amount of time with rate $\hat{\mu}_{ij}$ and assume (for tractability) that the physician can preempt an interaction to see a patient of a different class. When a physician returns to a preempted interaction, we assume s/he must repeat the process (e.g., review vital signs, lab results, etc.), and so we assume a preempt-repeat protocol.(In practice, emergency physicians can, and sometimes do, preempt patients to deal with emergencies. But for fairness and efficiency reasons, they do this rarely. Hence, we test our conclusions under the assumption of non-preemption in Phase 2 in Section 6 using a realistic simulation model.)

After each completed interaction, a patient of class ij may be disposed (discharged home or admitted to the hospital) with probability $p_{ij} > 0$, or with probability $1 - p_{ij}$ requires another round of test and treatment. We note that in practice the probability of being disposed may not be constant because it depends on various factors (e.g., progression of pain, the number of past interactions with the physician, revealed test

results, etc.). If data on such factors were collected, it could be incorporated into the patient prioritization decision. Since such data do not currently exist, we approximate the number of interactions with the physician by fitting a geometric distribution with constant probability of departure p_{ij} for class $ij \in U \times C$. Furthermore, we model test times, which include any preparation and wait times associated with the test, as a $\cdot/M/\infty$ queueing system with average service time of η^{-1}. Because we aggregate test times, waiting times for the test results, and preparations for tests into a single "test" stage, and also aggregate these for all possible types of tests, the long-run average time spent for a generic "test," denoted by η^{-1}, can be assumed to be roughly similar among different patient classes (for more detailed data on test turnaround times see [138] and [67]).

Because each physician is dedicated to his/her own slate of patients, we focus on a single physician's decision of who to see next. To this end, we let $\underline{x} = (x_{ij})_{ij \in U \times C}$ (respectively $\underline{y} = (y_{ij})_{ij \in U \times C}$) represent the number of patients of each class available (not available) for the physician visit. With these, we can define the state of the system at any point of time, t, by the vector $(\underline{x}(t), \underline{y}(t)) \in \mathbb{Z}_+^4 \times \mathbb{Z}_+^4$, and model the process $\{(\underline{x}(t), \underline{y}(t)) : t \geq 0\}$ as a Continuous Time Markov Chain (CTMC). We assume the parameters of the system are such that this CTMC is stablizable; i.e., there exists at least one policy under which the risk of adverse events is finite (otherwise, the problem does not represent a real ED). However, notice that since the transition rates are not bounded, we cannot use uniformization in the spirit of [99] to formulate a discrete time equivalent of the CTMC where the times between consecutive events are i.i.d. (for all states). However, in what follows, we construct a sequence of Controlled CTMC's (CCTMC's) with an increasing but bounded sequence of (maximum) transition rates converging to the original CCTMC. We do this by replacing the $\cdot/M/\infty$ test stage with four parallel $\cdot/M/k$ systems (one devoted to each patient class), index the underlying CCTMC with k, and let $k \to \infty$. The advantage of having four parallel $\cdot/M/k$ queues (instead of one $\cdot/M/k$) is that the order of

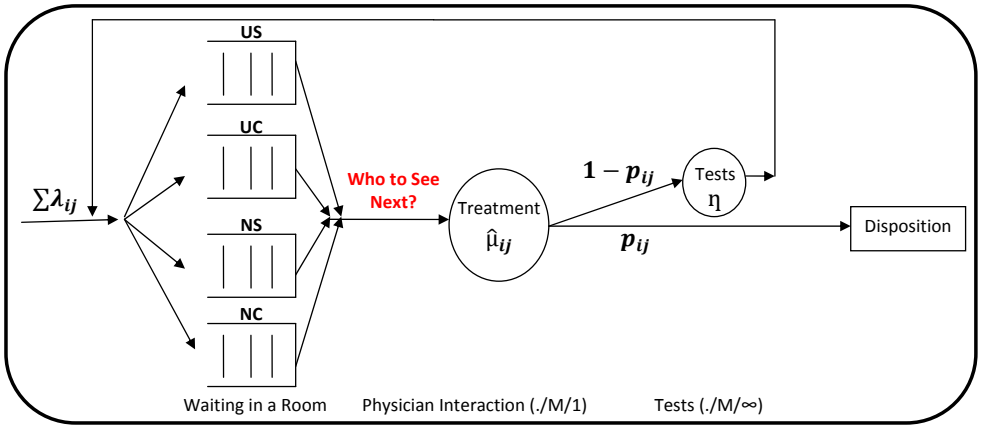


Figure 3.4: Patient flow after a patient is moved to an examination room/bed (Phase 2 sequencing).

jobs in each queue becomes irrelevant, and hence, does not need to be captured in the system's state. Another novel aspect of our approach is that we truncate the transition rates as opposed to the state space, thereby avoiding the artificial boundary effects that usually impact the policy. Since the transition rates in the CTMC indexed by k (for all k) are bounded by $\psi_k = \max_{ij \in U \times C} \hat{\mu}_{ij} + 4\,k\,\eta + \sum_{ij \in U \times C} \lambda_{ij} < \infty$, we can use the standard uniformization technique to derive the optimal policy for each CCTMC. We then use a convergence argument (taking the limit as $k \to \infty$) to derive the optimal policy for the original problem. It should be noted that we can always start with a sufficiently large k such that the stability of the underlying system is not affected (since the original system is stable by assumption).

For the system indexed by k, the optimal rate of adverse events, R^{k*}, and the optimal physician behavior can be derived from the following average cost optimality equation: $J^k(\underline{x}, \underline{y}) + R^{k*} =$

$$\frac{1}{\psi_k}\Big[\hat{\underline{\theta}}\,(\underline{x}+\underline{y})^T + \sum_{ij\in U\times C}[\lambda_{ij}J^k(\underline{x}+\underline{e}_{ij},\underline{y}) + (y_{ij}\wedge k)\,\eta\,J^k(\underline{x}+\underline{e}_{ij},\underline{y}-\underline{e}_{ij})]$$
$$+\min_{a\in\mathcal{A}(\underline{x})}\Big\{\sum_{ij\in U\times C}\mathbb{1}_{\{a=ij\}}\hat{\mu}_{ij}\,[p_{ij}\,J^k(\underline{x}-\underline{e}_{ij},\underline{y}) + (1-p_{ij})J^k(\underline{x}-\underline{e}_{ij},\underline{y}+\underline{e}_{ij})]$$
$$+\big(\psi_k - \sum_{ij\in U\times C}[\lambda_{ij}+(y_{ij}\wedge k)\,\eta + \mathbb{1}_{\{a=ij\}}\hat{\mu}_{ij}]\big)\,J^k(\underline{x},\underline{y})\Big\}\Big], \tag{3.7}$$

where $J^k(\underline{x},\underline{y})$ is a relative cost function (defined as the difference between the total expected cost of starting from state $(\underline{x},\underline{y})$ and that from an arbitrary state such as $(\underline{0},\underline{0})$), $a\wedge b=\min\{a,b\}$, $\underline{e}_{ij}$ is a vector with the same size as $\underline{x}$ with a 1 in position ij and zeroes elsewhere, a is an action determining which patient class to serve, and $\mathcal{A}(\underline{x})=\{ij\in U\times C: x_{ij}>0\}\cup\{0\}$ is the set of feasible actions (class 0 represents the idling action) when the number of patients of each class in the examination rooms is $\underline{x}$.

The optimal behavior of the physician is an appealing and simple operational rule, supporting implementation in practice.

Theorem 1 (Phase 2 Prioritization) *The physician should not idle when there is a patient available in an exam room. Furthermore, regardless of the number and class of available and unavailable patients, the physician should prioritize available patients in decreasing order of* $p_{ij}\,\hat{\theta}_{ij}\,\hat{\mu}_{ij}$.

Theorem 1 provides a simple prioritization index for physicians computed as the probability that the visit will be the final interaction with the patient (p_{ij}) times the estimated risk of adverse events ($\hat{\theta}_{ij}$) divided by the average duration of each visit ($1/\hat{\mu}_{ij}$). Such a policy is easy to implement, since (a) the physician does not need to consider the number and class of patients available in the examination rooms or under tests, and (b) the physician (or a decision support system) can dynamically estimate the required quantities. The authors have developed a smart phone application that

can be used by physicians to facilitate collection of required data and computation of patient priorities.

The above analysis confirms our intuition that a simple priority rule for Phase 2 is optimal. Moreover, the Phase 2 priority rule is consistent with that of Phase 1, since $1/\mu_{ij} = 1/(p_{ij}\hat{\mu}_{ij})$.

3.6 A Realistic Simulation Analysis of Complexity-Based Triage

Our analytical models of the previous sections suggest that adding patient complexity to the triage process and using appropriate priority rules can improve the ED performance in terms of both patient safety (ROAE) and operational efficiency (LOS). Furthermore, they provide some insights into hospital conditions under which such improvements are more beneficial. In this section, we test the conjectures suggested by our analytic models by means of a detailed ED simulation model. This simulation incorporates many realistic features of the University of Michigan ED (UMED) that are representative of most ED's in large research hospitals, including dynamic non-stationary arrivals, multi-stage service, multiple physicians and exam rooms, inaccuracy in triage classifications (both in terms of urgency and complexity), and limits on the number of patients physicians handle simultaneously. Our base case model uses a year of data from UMED plus time study data from the literature. We first describe the main features of our simulation framework, and then describe the test cases and our conclusions from them.

Patient Classes. At the time of triage, patients are classified according to both urgency (urgent or non-urgent) and complexity (simple or complex). For modeling purposes, we omit the resuscitation unit (RU) and fast track (FT) classifications, shown in Figure 3.1 (right), since these patients are typically tracked separately from the main ED. Following the definition of complex patients in [153], we define S patients

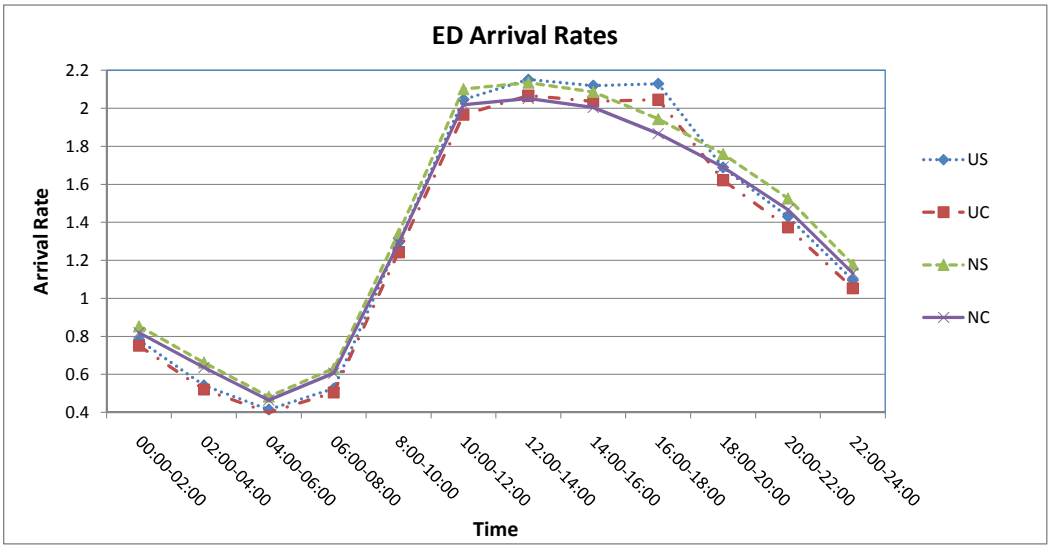


Figure 3.5: Class dependent arrival rates to the ED for an average day (obtained from a year of data in UMED).

as those who only require one treatment related interaction and C patients as those requiring two or more treatment related interactions.(To clarify, we do not count social interactions as a treatment related interaction. Furthermore, we would still classify a case as simple if the physician were first to order a test (without spending time treating the patient) and after receiving the results, conduct one treatment visit prior to discharge.) With ESI-4 and 5 patients omitted, we can equate U patients with ESI-2 patients, and N patients with ESI-3 patients for our purposes. Since the majority (about 80% in university of Michigan Hospital ED) of ED patients are composed of ESI-2 and ESI-3 patients, improvements for this subset of patients will have a major impact on overall ED performance. Both urgency and complexity classifications at the point of triage are subject to errors with different error rates. We assume the true type of a patient is not known until the final disposition decision is made. Consistent with the empirical findings of [63] and [153], we assume urgency and complexity classifications are subject to 10% and 17% error rates, respectively. For simplicity, we also assume urgency-based and complexity-based misclassification rates are independent and symmetric (i.e., triage nurses are equally likely to classify

U (C) patients as N (S) as they are to classify N (S) patients as U (C), respectively). But we consider asymmetric errors in our sensitivity analysis.

Arrival Process. Class-based patient arrivals are modeled using non-stationary Poisson processes that approximate our data. The non-stationary arrival rates for different classes are depicted in Figure 3.5. These arrival rates were obtained from a year of UMED data using the ESI levels based on two-hour intervals of the day. However, since patients are not currently triaged based on complexity, we used the empirical results of [153] (who found that about 49% of patients are complex) to obtain these arrival rates using a (stationary) splitting mechanism. The resulting pattern illustrated in Figure 3.5 is similar to those reported in other studies (e.g., [54]). A "thinning" mechanism (see [96] and [97]) is used to simulate the non-stationary Poisson process arrivals for each class of patients (with rates depicted in Figure 3.5) in our base case.

Service Process. The ED service process has multiple stages as depicted in the schematic in Figure 3.4. Each patient goes through one or more phases of patient-physician interactions followed by test/preparation/wait activities during which the physician cannot have a direct interaction with the patient (all such stages are labeled as Test in Figure 3.4). We also consider the initial and final preparations by a nurse. The initial preparation happens when the patient is moved to an exam room for the first time (before the first interaction with the physician) and the final preparation happens after the final visit by the physician and before the patient is discharged home or admitted to the hospital. The duration of each physician interaction is random and its average may depend on the class of the patient and the number of previous interactions. Our data suggest that the first and last interactions are typically longer than the intermediate interactions. As mentioned before and illustrated in Figure 3.1 (right), S patients are defined to be those who have only one (treatment related) interaction. For C patients, we can estimate the distribution of the number of physician interactions per patient as shown in Figure 3.6 using data from a detailed time study

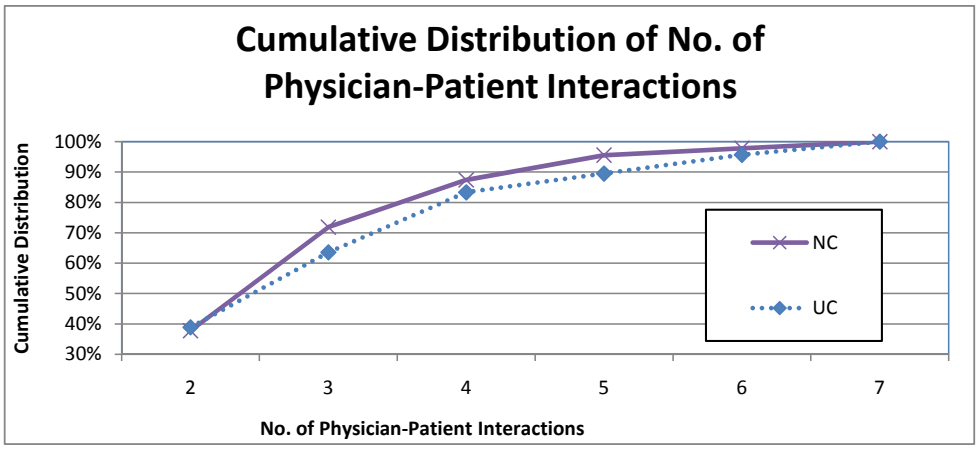


Figure 3.6: Cumulative number of class-based physician-patient interactions for complex patients (those requiring more than one interaction).

(see Table 3 of [52]) (normalized to represent our NC and UC patient classes). The simulated service process is considered to be non-collaborative, since an ED physician rarely transfers his/her patients to another physician, and also non-preemptive.

Physician-Patient Assignments and Priorities. As mentioned earlier, the process of connecting patients with physicians involves two phases. In Phase 1, patients are brought back from the waiting area to exam rooms whenever a room becomes available based on a Phase 1 sequencing priority. Phase 1 is usually performed by a charge nurse. In Phase 2, whenever a physician becomes available, and if s/he has fewer than his/her maximum number of patients (7 is typical), s/he chooses the next patient from those available based on a Phase 2 sequencing rule, which will depend on the type of triage being used. For urgency-based triage, we assume U patients get priority over N patients in both Phases 1 and 2. For complexity-based triage, patients are prioritized in both Phases according to the strict priority ordering US, UC, NS, NC (ranked from high to low priority) which we found to be optimal in the simplified ED models discussed previously (see Proposition 5). When a patient is brought back to an examination room, we assume that s/he is assigned to the physician with the lowest number of patients. If all physicians are handling more than 7 patients, the

patient must wait. Phase 1 and Phase 2 priority decisions can only be made based on the estimated class of the patient, which is subject to misclassification error, but adverse events are determined by the true class of the patient.

ED Resources. We consider 22 beds and 4 physicians in our base case scenario. We then perform a sensitivity analysis to understand the effect of number of both beds and physicians on the benefit of complexity-based triage. For simplicity, we do not consider end of shift effects and/or variations in the level of staff available. Furthermore, we consider test facilities (ancillary services) as exogenous resources (i.e., test times are independent of the volume of ED patients) because these facilities often handle many other patients besides those from the ED.

Adverse Events. Adverse events are simulated using Poisson processes with rates that depend on the class of patients, as well as the phase of service. Specifically, we assume that U patients have a higher rate of adverse events than N patients, and that after patients enter an exam room (Phase 2 of service), their rate of adverse events decreases by 60% (in our base case) relative to their rates in the waiting area (Phase 1 of service). As in our previous models, we do not consider fatal events that would terminate the adverse events counting process, since the impact of these rare events on our objective function is extremely small.

Runs. The simulation was written in a C++ framework and makes use of a cyclo-stationary model with a period of a week. Each data point was obtained for 5000 replications of one week, where each replication was preceded by a warm-up period of one week (which was observed to be sufficient because correlations in the ED flow are very small for spans of two or more days). The number of replications (5000) was chosen so that the confidence intervals are tight enough that (1) the sample averages are reliable, and (2) we can omit these very tight intervals from our data presentations.

In the following sections, we describe how we used our simulation model to analyze the benefit of complexity-based triage over urgency-based triage.

3.6.1 Performance of Complexity-Based Triage

We start by comparing complexity-based triage to urgency-based triage in our base case model, under the assumption that both types of triage make use of their respective priority rules for sequencing patients in both Phase 1 and Phases 2. This leads to the following:

Observation 1. *In the base case, implementing complexity-based triage improves ROAE and LOS by* 9.41% *(0.16 events/hr) and* 7.68% *(36 mins/patient), respectively.*

To consider the case where Phase 2 sequencing cannot follow the optimal rule due to a lack of data, patient discomfort, or other factors, we also compare complexity-based triage with urgency-based triage when Phase 2 sequencing in both systems uses a service-in-random-order (SIRO) rule. This leads to improvements of 7.95% and 7.01% in ROAE and LOS, respectively. Hence, it appears that the benefits of complexity-based triage are robust to the policy used in Phase 2. At least in our base case, it is the refined sequencing in Phase 1 that drives the majority of the improvement.

The smaller effect of Phase 2 sequencing compared to that of Phase 1 prioritization is mainly due to the fact that, under the conditions of our base case, physicians in Phase 2 often do not have many available patients from which to choose. This is because (a) patients are unavailable for a considerable amounts of time while being tested and waiting for test results, and (b) each physicians is handling a limited number of patients at a time (with a constrained upper bound of seven). However, in ED's with shorter test times (e.g., more test facilities dedicated to the ED, or more responsive central test facilities), larger case loads (patients per physician), and enough examination rooms/beds to accommodate patients, there will be more choices among in-process patients, and hence more improvement from an effective Phase 2 sequencing policy. To test this, we consider an ED with test rates 70% faster than the base case values, 40 beds, 3 physicians, and a maximum number of 10 patients per physician. Under these conditions, if Phase 2 sequencing is done according to SIRO for

both the urgency-based and complexity-based triage systems, then complexity-based triage achieves improvements of 8.58% and 6.15% in ROAE and LOS, respectively, relative to urgency-based triage. In contrast, if the urgency-based triage system prioritizes patients in Phase 2 by urgency ($U > N$) and the complexity-based triage system prioritizes patients in Phase 2 by complexity and urgency ($US > UC > NS > NC$), then complexity-based triage achieves improvements of 13.09% and 9.11% in ROAE and LOS, respectively, relative to urgency-based triage. This leads us to the following:

Observation 2. *In ED's where physicians have more choice about what patient to see next, using complexity information to prioritize patients in Phase 2 becomes more valuable.*

3.6.2 How to Define Complex Patients?

In the previous section, we investigated the benefit of complexity-based triage using the approach of [153] to define complex patients as those requiring at least two (treatment related) interactions with a physician. This results in a nearly even split between complex and simple patients (49% C vs. 51% S), as well as substantial heterogeneity between their treatment time (both of which were predicted in Proposition 6 to be factors that improve the performance of complexity-based triage). But we could use other standards for defining a patient to be complex. In Figure 3.7, we illustrate the impact of complexity-based triage on ROAE and LOS when complex patients are defined to be as those with more than one (resulting in 49% C patients), more than two (resulting in 39% C patients), and more than three (resulting in 30% C patients) interactions. From this we conclude:

Observation 3. *If the number of (treatment related) interactions is used as the metric for patient complexity, the benefit of complexity-based triage is greatest when complex patients are defined to be those requiring at least two interactions.*

The reason for this is that increasing the number of interactions required for a

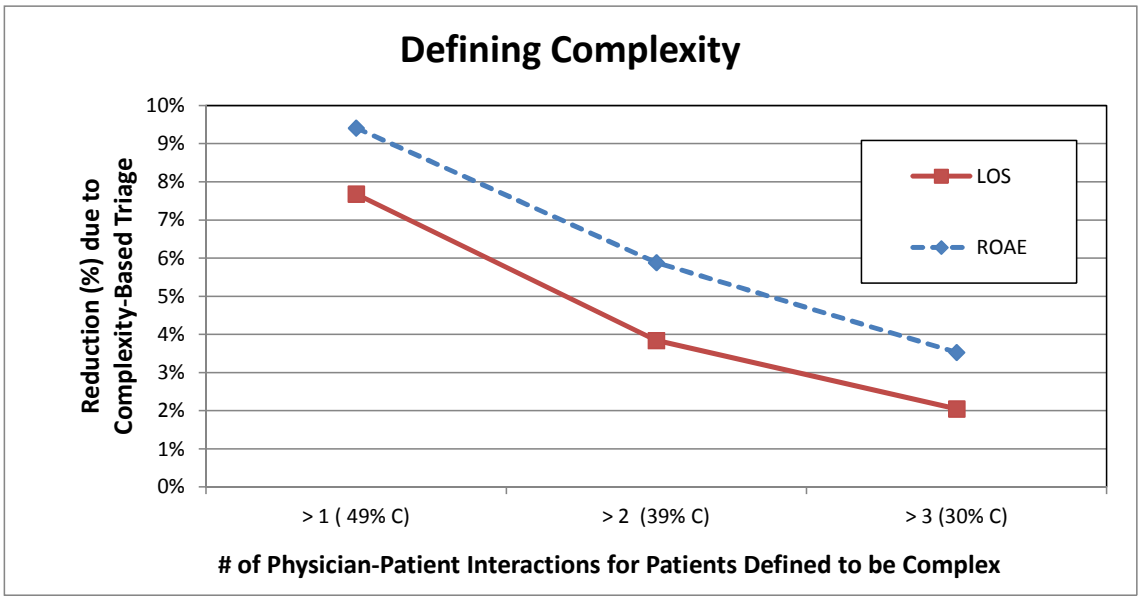
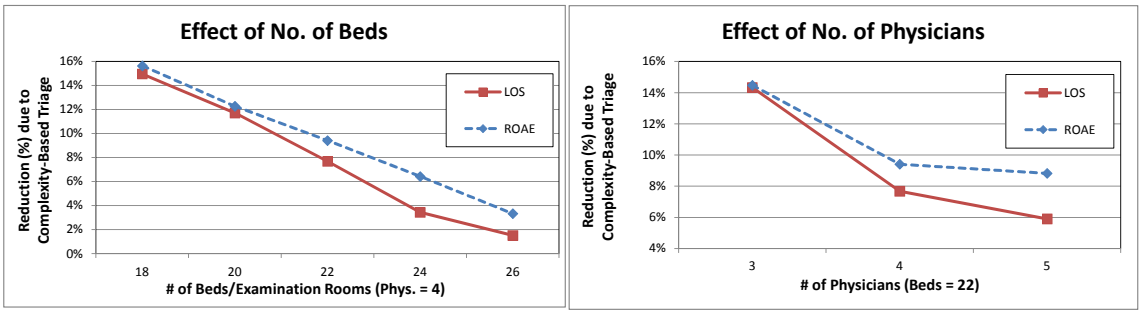


Figure 3.7: Performance of complexity-based triage when defining complex patients to be those having more than one, more than two, and more than three physician-patient interactions.

patient to be considered complex decreases the fraction of complex patients substantially, but only slightly increases the difference in treatment times between complex and simple patients. Thus, as predicted by Proposition 6, the benefit of complexity-based triage declines.

3.6.3 The Effect of ED Resource Levels

Another factor predicted by Proposition 6 to favor complexity-based triage is resource utilization. In that proposition, resources refer to physicians and examination rooms (which are indistinguishable in the single-stage simplified ED model). Hence, we expect higher utilization of either physicians or examination rooms to increase the benefit of complexity-based triage. Figure 3.8 illustrates the percentage improvement (in terms of ROAE and LOS) of complexity-based triage over urgency-based triage for varying numbers of examination rooms and physicians. In addition to the LOS for patient classes considered (i.e., ESI 2 and 3) with 4 physicians, this figure also



Triage Policy	Phys. = 4, Beds = 18		Phys. = 4, Beds = 20		Phys. = 4, Beds = 22		Phys. = 4, Beds = 24		Phys. = 4, Beds = 26	
	TTR (hrs)	LOS (hrs)	TTR (hrs)	LOS (hrs)	TTR (hrs)	LOS (hrs)	TTR (hrs)	LOS (hrs)	TTR (hrs)	LOS (hrs)
Urgency-Based	31.57	35.24	10.66	14.43	3.91	7.81	1.82	5.81	1.16	5.25
Complexity-Based	26.26	29.97	8.95	12.74	3.3	7.21	1.59	5.61	1.04	5.17

Figure 3.8: The effect of resources (beds and physicians) on the benefit of complexity-based triage over the current practice of urgency-based triage [Left: the effect of beds (4 physicians); Right: the effect of physicians (22 beds)].

presents the average time spent in Phase 1, labeled as Time to Room (TTR), under each triage system. From this figure we observe the following:

Observation 4. *The benefit of complexity-based triage is greater in ED's with higher bed and/or physician utilization.*

As we observed in the Introduction, most ED's are overcrowded, so high utilization is a common situation. Hence, results from our analytic and simulation models suggest that complexity-based triage is most effective precisely in ED's most in need of improvement.

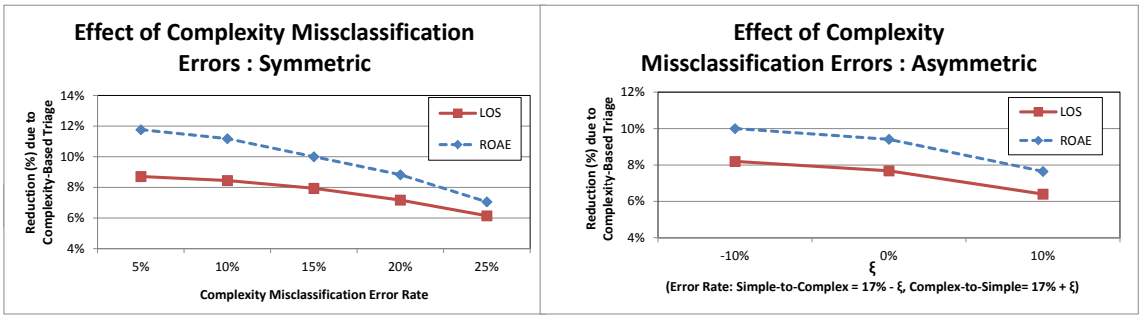


Figure 3.9: The effect of complexity misclassification error rates on the benefit of a complexity-based triage (compared to an urgency-based only) [Left: symmetric misclassification; Right: asymmetric misclassification].)

3.6.4 The Effect of Misclassification

Finally, we investigate the impact of complexity-based misclassification errors, which are inevitable in any triage system. Figure 3.9 (left) shows the benefits (in ROAE and LOS) of complexity-based triage over urgency-based triage for variations of the base case, in which complexity misclassification error rates range from 5% to 25%. Figure 3.9 (left) assumes these errors to be symmetric; that is, the chance of classifying an S patient as C is equal to the chance of classifying a C patient as S. Figure 3.9 (right) considers asymmetric error rates while keeping the average misclassification rate constant and equal to the base-case value of 17% (reported in the empirical study of ([153]). From these figures, we observe the following:

Observation 5. *The benefit of complexity-based triage is robust to complexity misclassification errors. However, complex-to-simple misclassifications are slightly more harmful than simple-to-complex misclassifications.*

The intuition behind the second part of this observation is that a complex-to-simple misclassification error moves a complex patient up in the queue, potentially delaying many other patients. In contrast, a simple-to-complex misclassification error moves a simple patient back in the queue, delaying only that patient. So, it is slightly better to err on the side of classifying ambiguous patients as complex rather than simple.

3.6.5 Complexity-Based Streaming

In the previous sections, we investigated the benefit of collecting and using complexity-based information to prioritize patients in the ED. But another way to make use of this information is to separate patients into different service streams for simple and complex patients (somewhat analogous to the admit/discharge streaming implemented in Flinders Medical Center ([90]) with complexity information used in place of admit/discharge predictions). We are interested in whether such streaming is more

effective than pooling-based prioritization.

To investigate this design question we raised in the Introduction, we examine a *complexity-based streaming* patient flow design in which two service streams of patients are created: one for patients triaged as simple (S) and one for those triaged as complex (C). The resources (beds and physicians) are labeled with S and C, indicating their main purpose. However, to overcomes the "anti-pooling" disadvantage of streaming, we allow the resources to be assigned to the other stream as needed, which is a feature we found to be useful in [121]. For instance, when a C physician is available but there is no complex patient available, the physician can be assigned to an S patient who is waiting. In this design, we assume that patients in each stream and in both Phases 1 and 2 are prioritized according to their ESI level.

Since simple and complex patients are separated, lean process improvement techniques can be implemented to improve and standardize service, particularly on the simple side for which the repetitive treatment processes can be organized in a clear, flow-shop like path. In Figure 3.10, we compare the performance of the complexity-based streaming design, with and without such lean improvements, against that of urgency-based pooling (current practice) and complexity-based pooling (i.e., a pooling design where Phase 1 and Phase 2 are based on the optimal priority rule using complexity-based triage information). The system with lean implementation assumes the service rate for each interaction with the simple patients improves by 10%; however, no change occurs for complex patients. This is a conservative estimate of the impact of a lean transformation. Note that the streaming layout facilitates this improvement by grouping simple tasks in a single line. Under pooled designs the mixture of simple and complex patients makes a smooth efficient flow extremely difficult. It should be also noted that some lean improvements may be possible for complex stream, but we conservatively ignore that here. Figure 3.10 compares performance in terms of LOS, but we have observed similar results for the ROAE criterion. These results lead to the following:

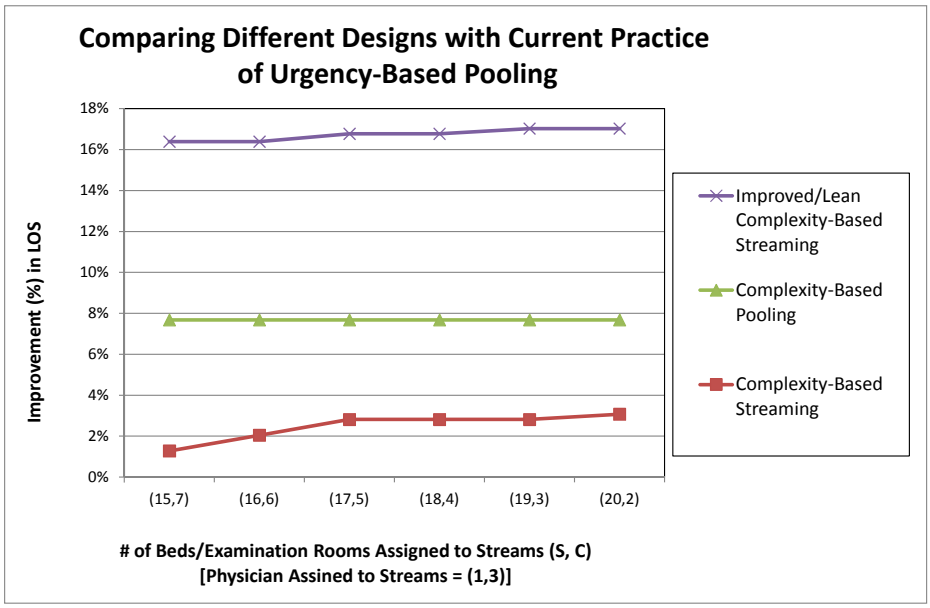


Figure 3.10: Performance of different patient flow designs compared to the current practice (Urgency-Based Pooling).

Observation 6. *Without lean improvements, complexity-based streaming is still better than current pooling practice, but worse than complexity-based pooling. With lean improvements (made only to the simple stream), complexity-based streaming can achieve a substantial advantage over complexity-based pooling.*

3.7 Conclusion

In this chapter, we propose a new triage system for ED practice in which patients are classified on the basis of complexity, as well as urgency. Our results suggest that, compared to the triage system currently in use in practice, complexity-based triage can significantly improve ED performance in terms of both patient safety (ROAE) and operational efficiency (LOS), even if patient classification is subject to error.

We also investigate effective but implementable policies for prioritizing patients in the ED. We show that the current practice of prioritizing patients purely based on urgency (e.g., ESI 2 over 3 in the main ED) is suboptimal, and it is essential to

take into account a measure of patient complexity. This can address many of the performance limitations of the current triage system in ED's that are widely reported by clinicians (see, e.g., [160] and the references therein).

We find that a simple and fast classification scheme, which defines patients to be simple if they require only a single treatment related interaction (and complex otherwise) works very well as the basis for complexity-based triage as it results in (1) a nearly even split between simple and complex patients, and (2) a substantial difference between average treatment time of complex and simple patients. This classification scheme has been empirically shown (see [153]) to be feasible for nurses to implement at triage with reasonable accuracy, and hence, appears to be a promising enhancement of the triage process.

To accomplish this research, we developed new models, contributed several analytical contributions, collected hospital data, and developed high-fidelity simulations. We advanced the analysis of priority queueing systems under misclassification errors as well as continuous-time MDP analysis with unbounded transition rates, for which the traditional method of uniformization fails. Using these technical innovations, we show that new, extended versions of the $c\mu$ rule can provide effective guidelines for prioritizing patients in both Phase 1 and Phase 2 of service in the ED, even when many practical conditions in the ED are considered.

Our analyses indicate that complexity-based triage can yield substantial safety and efficiency improvements even if complexity information is only used to prioritize patients up to the point where they enter examination rooms (Phase 1). Furthermore, in ED's where physicians have a significant amount of choice about what patient to see next within examination rooms (Phase 2), we find that complexity information gathered at triage can yield additional benefits by facilitating internal sequencing decisions. For both Phase 1 and Phase 2, the benefit of complexity-based triage is greatest in ED's with high physician and/or examination room utilization. Since ED's are widely overcrowded, our results suggest that complexity-based triage is

an effective way for ED's to improve safety and reduce congestion without adding expensive human or physical capacity.

We further investigate a new patient flow design, in which complexity-based triage information is used to separate simple and complex patients into two streams. Our results show that, when combined with improvements achieved through implementation of lean methods on the "simple" patient service stream, this complexity-based streaming design can take advantage of complexity-based triage information to achieve even greater gains.

3.8 Appendix (Proofs)

Proof of Proposition 4: The proof of part (i) follows directly from comparing (3.4) and (3.5). To show part (ii), notice that, using the result of part (i) for a special case where there is no misclassification error, prioritizing U (N) patients is optimal if, and only if, $\theta_U\,\mu_U \geq (\leq)\theta_N\,\mu_N$. Next, observe that $\theta'_U\,\mu'_U - \theta'_N\,\mu'_N = \big[\lambda_N\lambda_U\mu_N\mu_U(\theta_U\mu_U - \theta_N\mu_N)(1-\gamma_N-\gamma_U)\big]/\big[\big(\lambda_N\mu_U\gamma_N + \lambda_U\mu_N(1-\gamma_U)\big)\big(\lambda_N\mu_U(1-\gamma_N) + \lambda_U\mu_N\gamma_U\big)\big]$. Combining these two results completes the proof of part (ii), as the sign of the numerator changes when the sum of errors exceeds 1. □

Lemma 2 (Perfect Classification - Prioritization) *In the simplified single-stage ED model under perfect urgency and complexity based classification:*
(i) The best priority rule is to prioritize patients in decreasing order of $\theta\mu$ values. Hence, if $\theta_{UC}\mu_{UC} \geq \theta_{NS}\mu_{NS}$, then the best priority rule is to follow the ordering: US, UC, NS, NC. Otherwise, the ED should follow the priority ordering: US, NS, UC, NC.
(ii) $R_*^{U\cup C} \leq R_*^U$. That is, the risk of adverse events under the optimal priority rule using both complexity and urgency information is (weakly) smaller than that under the optimal apriority rule using only urgency information.

(iii) The best priority rule of part (i) is optimal even among the larger class of all non-anticipative policies (state or history dependent, idling or non-idling, etc.).

Proof of Lemma 2: Notice that, using (3.1), we can compute the average waiting time of each class of patients under any (static) priority rule. Furthermore, under priority rule π, we have

$$R_{\pi}^{U\cup C} = \sum_{i\in U}\sum_{j\in C}\theta_{ij}\lambda_{ij}W_{ij}^{\pi}, \tag{3.8}$$

where W_{ij}^{π} is the average waiting of class ij under priority rule π. The proof of part (i) then follows from [31] (see pages 83-84), where an interchange argument is used (when the number of customer classes is at least 3) to show that the best rule (among the priority policies) to minimize the holding cost in a non-preemptive M/G/1 is to follow the $c\mu$ rule. Replacing holding cost values (c) with adverse event rates (θ), and noticing that the patient class US (NC) has the highest (lowest) $\theta\mu$ value complete the proof of part (i). Next, using the result of part (i) together with (3.1) and (3.8), when $\theta_{UC}\mu_{UC} \geq \theta_{NS}\mu_{NS}$, we have:

$$\begin{aligned} R_{*}^{U\cup C} &= \lambda\mathbb{E}(s^2)\Big[\frac{\lambda_{US}\theta_{US}}{2(1-\rho_{US})} + \frac{\lambda_{UC}\theta_{UC}}{2(1-\rho_{US})(1-\rho_{US}-\rho_{UC})} \\ &\qquad + \frac{\lambda_{NS}\theta_{NS}}{2(1-\rho_{US}-\rho_{UC})(1-\rho_{US}-\rho_{UC}-\rho_{NS})} \\ &\qquad + \frac{\lambda_{NC}\theta_{NC}}{2(1-\rho_{US}-\rho_{UC}-\rho_{NS})(1-\rho_{US}-\rho_{UC}-\rho_{NS}-\rho_{NC})}\Big] \qquad (3.9) \\ &\leq \min\{R_{U}^{U}, R_{N}^{U}\} = R_{*}^{U}, \end{aligned}$$

where the inequality follows from (3.2) and (3.3) together with the result of part (i) of Proposition 4 (for the special case where there is no misclassification error).

When $\theta_{UC}\mu_{UC} < \theta_{NS}\mu_{NS}$, we have:

$$R_*^{U\cup C} = \lambda\mathbb{E}(s^2)\Big[\frac{\lambda_{US}\theta_{US}}{2(1-\rho_{US})} + \frac{\lambda_{NS}\theta_{NS}}{2(1-\rho_{US})(1-\rho_{US}-\rho_{NS})} + \frac{\lambda_{UC}\theta_{UC}}{2(1-\rho_{US}-\rho_{NS})(1-\rho_{US}-\rho_{NS}-\rho_{UC})} + \frac{\lambda_{NC}\theta_{NC}}{2(1-\rho_{US}-\rho_{UC}-\rho_{NS})(1-\rho_{US}-\rho_{UC}-\rho_{NS}-\rho_{NC})}\Big], \tag{3.10}$$

and similar to the previous case, it can be easily seen that $R_*^{U\cup C} \leq R_*^U$. The proof of part (iii) follows from [86] (after replacing holding cost with adverse event rates) who (for the average holding cost objective) showed that the $c\mu$ policy of [31] remains optimal even when inserting idleness is allowed and/or when the priority rule is dynamic (i.e., state-dependent). □

Lemma 3 (Perfect Classification - Attractiveness) *In the simplified single-stage ED model, perfect complexity-based triage yields a larger improvement over perfect urgency-based triage when (i) ED utilization is higher, (ii) heterogeneity in the average service time of simple vs. complex patients is larger, and/or (iii) the fraction of simple and complex patients are closer to equal.*

Proof of Lemma 3: To show the result, first consider the case where under the $U\cup C$ classification it is optimal to follow the priority order US, UC, NS, NC, and under the U classification, it is optimal to follow the priority order U, N (i.e., prioritizing urgent patients first). Let $f = R_*^{U\cup C} - R_*^U$, and notice that with $\mu_{iC} = \mu_C$ and $\mu_{iS} = \mu_S$ ($\forall i \in U$), and $\theta_{Uj} = \theta_U$ and $\theta_{Nj} = \theta_N$ ($\forall j \in C$) (i.e., when complexity is based only on set C and urgency is based only on set U), from (3.9) and (3.2) we have:

$$f = -[\frac{\theta_U\lambda_{US}\lambda_{UC}(1/\mu_C - 1/\mu_S)}{2(1-\rho_U)} + \frac{\theta_N\lambda_{NC}\lambda_{NS}(1/\mu_C - 1/\mu_S)}{2(1-\rho_U)(1-\rho)}]. \tag{3.11}$$

Then, a careful treatment of utilization (realizing that $\rho_U = \lambda_U/\mu_U$ and $\rho = \rho_U + \rho_N$) shows that f is non-increasing in utilization, ρ. To prove part (ii), it then can be seen that f is non-increasing in $1/\mu_C - 1/\mu_S$ (keeping utilization and other factors the same). To see part (iii), let $\alpha \in [0, 1]$ denote the fraction of patients that are complex, and $(1-\alpha)$ denote the fraction of patients that are simple, so $\lambda_{US} = (1-\alpha)\lambda_U$, $\lambda_{UC} = \alpha\lambda_U$, $\lambda_{NC} = \alpha\lambda_N$, and $\lambda_{NS} = (1-\alpha)\lambda_N$. Replacing these in (3.11), it follows that f, as a function of α, can be written as $f = -[\alpha(1-\alpha)]k$, for some constant $k \geq 0$. Thus, $\alpha = 0.5$ yields the maximum benefit. The proof for other cases (i.e., when other priority rules are optimal) follows a similar argument after computing f using either (3.9) or (3.10), and either (3.2) or (3.3), depending on the optimal priority rule under $U \cup C$ and U classifications, respectively. □

Proof of Proposition 5: The proof of part (i) follows directly from the proof of part (i) of Lemma 2, since all rates are replaced with their error impacted counterparts. That is, the same interchange method of [31] (see pages 83-84) after replacing all rates with their error impacted counterparts proves that the best priority rule is to give priority based on a decreasing order of $\theta\mu$ values. The proof of part (ii) follows from the proof of Lemma 2 (found earlier in this appendix) part (ii) after replacing parameters with their error impacted counter parts. The proof of part (iii) follows from the result of [86], after replacing holding cost with the error impacted rate of adverse events, and all the other rates with their error impacted counterparts. □

Proof of Proposition 6: The proof of parts (i) - (iii) follows mainly from the proof of Lemma 3 (found earlier in this appendix). First, consider the case where under the $U' \cup C'$ (i.e., imperfect urgency and complexity) classification it is optimal to follow the priority order US, UC, NS, NC, and under the U' (i.e., imperfect urgency) classification, it is optimal to follow the priority order U, N (i.e., prioritizing urgent patients over non-urgent patients). With $f = R_*^{U' \cup C'} - R_*^{U'}$, and after replacing rates

with their error impacted counterparts in (3.11) we have:

$$f = -[\frac{\theta'_U \lambda'_{US} \lambda'_{UC} (1/\mu'_C - 1/\mu'_S)}{2(1-\rho'_U)} + \frac{\theta'_N \lambda'_{NC} \lambda'_{NS} (1/\mu'_C - 1/\mu'_S)}{2(1-\rho'_U)(1-\rho')}]. \tag{3.12}$$

Next, notice that $\rho' = \rho$ (i.e., the total utilizations with and without misclassifications are the same). Hence, similar to the proof of part (i) of Lemma 3, it can bee seen that f is non-increasing in ρ. Moreover, it can be seen that f is non-increasing in $1/\mu'_C - 1/\mu'_S$. Next, notice that $(\underline{1}/\underline{\mu'})^T = (A\,(\underline{\lambda}/\underline{\mu})^T)/\underline{\lambda'}$, where A is defined in (3.6). Thus, similar to the proof of part (ii) of Lemma 3, it can be seen that f is non-increasing in $1/\mu_C - 1/\mu_S$, which proves part (ii). Furthermore, similarly to the proof of part (iii) of Lemma 3, let $\lambda_{US} = (1-\alpha)\lambda_U$, $\lambda_{UC} = \alpha\lambda_U$, $\lambda_{NC} = \alpha\lambda_N$, and $\lambda_{NS} = (1-\alpha)\lambda_N$. It can be seen that f as a function of α is minimized at $\alpha = 0.5$, which proves part (iii). It can also be seen that f is non-decreasing in complexity misclassification error rates, γ_S and γ_C, which proves part (iv). The proof for other cases (i.e., when other priority rules are optimal) follows a similar line of argument after computing f. □

Proof of Theorem 1: To show the result, we use an *interchange* argument; we show that if classes $uc \in U \times C$ and $sl \in U \times C$ are such that $p_{uc}\,\hat{\theta}_{uc}\,\hat{\mu}_{uc} \geq p_{sl}\,\hat{\theta}_{sl}\,\hat{\mu}_{sl}$, then it is (weakly) better to serve class uc than class sl when in state $(\underline{x}, \underline{y})$ with $x_{uc}, x_{sl} > 0$. This will also prove that the optimal policy will not idle the physician when there are one or more patients available in the rooms, since idling can be thought of serving an additional class, class 0, with $\hat{\theta}_0 = \hat{\mu}_0 = p_0 = 0$ (see, for instance, [22]). To show that it is (weakly) better to serve class uc than class sl, we first consider the problem in an N-period discounted cost setting with four parallel (one for each class of patients) $\cdot/M/k$ systems (to guarantee bounded transition rates for the purpose of uniformization) in place of the $\cdot/M/\infty$ test stage, and show that the results hold for any number of periods to go $n \in 1, 2, ..., N$. (Notice that using four parallel $\cdot/M/k$ systems removes the need for considering the sequence and the type of patients within

the common queue.) Using a convergence argument, as $n \to \infty$, it then follows that the result is true for an infinite-horizon (and hence, average cost) scenario with the four k-server test system. Next, taking limit as $k \to \infty$, it follows that the result is true even when transition rates are not bounded due to the existence of the $\cdot/M/\infty$ stage.

Now consider the finite horizon discounted cost version of (3.7). With β denoting the discount factor, the optimal discounted cost when there are $n+1$ (uniformized) periods to go is $V_{n+1}^k(\underline{x}, \underline{y}) =$

$$\frac{1}{\psi_k}\Big[\hat{\underline{\theta}}\,(\underline{x}+\underline{y})^T + \beta\big[\sum_{ij\in U\times C}[\lambda_{ij}V_n^k(\underline{x}+\underline{e}_{ij},\underline{y}) + (y_{ij}\wedge k)\,\eta\,V_n^k(\underline{x}+\underline{e}_{ij},\underline{y}-\underline{e}_{ij})]$$
$$+\min_{a\in\mathcal{A}(\underline{x})}\Big\{\sum_{ij\in U\times C}\mathbb{1}_{\{a=ij\}}\hat{\mu}_{ij}\,[p_{ij}\,V_n^k(\underline{x}-\underline{e}_{ij},\underline{y}) + (1-p_{ij})V_n^k(\underline{x}-\underline{e}_{ij},\underline{y}+\underline{e}_{ij})]$$
$$+\big(\psi_k - \sum_{ij\in U\times C}[\lambda_{ij}+(y_{ij}\wedge k)\,\eta + \mathbb{1}_{\{a=ij\}}\hat{\mu}_{ij}]\big)\,V_n^k(\underline{x},\underline{y})\Big\}\big]\Big], \tag{3.13}$$

or equivalently (grouping the terms related to control in the minimization and self-loop), $V_{n+1}^k(\underline{x}, \underline{y}) =$

$$\frac{1}{\psi_k}\Big[\hat{\underline{\theta}}\,(\underline{x}+\underline{y})^T + \beta\big[\sum_{ij\in U\times C}[\lambda_{ij}V_n^k(\underline{x}+\underline{e}_{ij},\underline{y}) + (y_{ij}\wedge k)\,\eta\,V_n^k(\underline{x}+\underline{e}_{ij},\underline{y}-\underline{e}_{ij})]$$
$$-\max_{a\in\mathcal{A}(\underline{x})}\Big\{\sum_{ij\in U\times C}\mathbb{1}_{\{a=ij\}}\hat{\mu}_{ij}\,[p_{ij}\Delta_{ij}^y\,V_n^k(\underline{x}-\underline{e}_{ij},\underline{y}) + \Delta_{ij}^{x,y}V_n^k(\underline{x}-\underline{e}_{ij},\underline{y}+\underline{e}_{ij})]$$
$$+\big(\psi_k - \sum_{ij\in U\times C}[\lambda_{ij}+(y_{ij}\wedge k)\,\eta]\big)\,V_n^k(\underline{x},\underline{y})\Big\}\big]\Big], \tag{3.14}$$

where $\Delta_{ij}^y\,V_n^k(\underline{x},\underline{y}) = V_n^k(\underline{x},\underline{y}+\underline{e}_{ij}) - V_n^k(\underline{x},\underline{y})$ and $\Delta_{ij}^{x,y}V_n^k(\underline{x},\underline{y}) = V_n^k(\underline{x}+\underline{e}_{ij},\underline{y}-\underline{e}_{ij}) - V_n^k(\underline{x},\underline{y})$. Now let π ($\hat{\pi}$) be the policy that prescribes serving patients of class uc (sl) for every state $(\underline{x},\underline{y})$ with $x_{uc}, x_{sl} > 0$ and in every period n. From (3.14), to show that π is (weakly) better than $\hat{\pi}$ in every period, we need to show that the

following property holds for every n and every state $(\underline{x}, \underline{y})$ with $x_{uc}, x_{sl} > 0$:

$$\begin{aligned}
&\hat{\mu}_{uc}\left[p_{uc}\Delta^{y}_{uc}\, V^{k,\pi}_{n}(\underline{x}-\underline{e}_{uc}, \underline{y}) + \Delta^{x,y}_{uc} V^{k,\pi}_{n}(\underline{x}-\underline{e}_{uc}, \underline{y}+\underline{e}_{uc})\right] \\
&\qquad \geq \hat{\mu}_{sl}\left[p_{sl}\Delta^{y}_{sl}\, V^{k,\hat{\pi}}_{n}(\underline{x}-\underline{e}_{sl}, \underline{y}) + \Delta^{x,y}_{sl} V^{k,\hat{\pi}}_{n}(\underline{x}-\underline{e}_{sl}, \underline{y}+\underline{e}_{sl})\right]. \qquad (3.15)
\end{aligned}$$

To show property (3.15), we use induction on n. First, for $n = 0$, the property trivially holds since $V^{\pi}_{0}(\cdot,\cdot) = V^{\hat{\pi}}_{0}(\cdot,\cdot) = 0$. Next, suppose the property holds for n. We show that it will then also hold for $n+1$. To do so, we need to consider different cases based on the state (i.e., partitions of the state space). First, consider the case where $x_{uc}, x_{sl} \geq 2$. Using action $a = uc$ (policy π) in both states $(\underline{x}-\underline{e}_{uc}, \underline{y})$ and $(\underline{x}, \underline{y})$ to compute $V^{k,\pi}_{n+1}(\underline{x}-\underline{e}_{uc}, \underline{y}+\underline{e}_{uc})$ and $V^{k,\pi}_{n+1}(\underline{x}-\underline{e}_{uc}, \underline{y})$ using (3.14), and subtracting the results we have $\Delta^{y}_{uc} V^{k,\pi}_{n+1}(\underline{x}-\underline{e}_{uc}, \underline{y}) =$

$$\begin{aligned}
\frac{1}{\psi_k}\Big[\hat{\theta}_{uc} + \beta\big[&\sum_{ij\in U\times C}\left[\lambda_{ij}\Delta^{y}_{uc}V^{k,\pi}_{n}(\underline{x},\underline{y}) + (y_{ij}\wedge k)\,\eta\,\Delta^{y}_{uc}V^{k,\pi}_{n}(\underline{x}+\underline{e}_{ij}-\underline{e}_{uc}, \underline{y}-\underline{e}_{ij})\right] \\
&\qquad + \mathbb{1}_{\{y_{uc}<k\}}\,\eta\, V^{k,\pi}_{n}(\underline{x}, \underline{y}-\underline{e}_{uc}) \\
&- \hat{\mu}_{uc}\left[p_{uc}\Delta^{y}_{uc}\Delta^{y}_{uc}\, V^{k,\pi}_{n}(\underline{x}-2\underline{e}_{uc}, \underline{y}) + \Delta^{y}_{uc}\Delta^{x,y}_{uc} V^{k,\pi}_{n}(\underline{x}-2\underline{e}_{uc}, \underline{y}+\underline{e}_{ij})\right] \\
&+ \big(\psi_k - \sum_{ij\in U\times C}\left[\lambda_{ij} + (y_{ij}\wedge k)\,\eta\right]\big)\,\Delta^{y}_{uc}V^{k}_{n}(\underline{x}-\underline{e}_{uc}, \underline{y}) \\
&\qquad - \mathbb{1}_{\{y_{uc}<k\}}\,\eta\, V^{k,\pi}_{n}(\underline{x}-\underline{e}_{uc}, \underline{y})\big]\Big]. \qquad (3.16)
\end{aligned}$$

Similarly, we can derive $\Delta^{x,y}_{uc} V^{k,\pi}_{n+1}(\underline{x}-\underline{e}_{uc}, \underline{y}+\underline{e}_{uc})$ using (3.14) and action $a = uc$ (policy π) in both states $(\underline{x}-\underline{e}_{uc}, \underline{y}+\underline{e}_{uc})$ and $(\underline{x}, \underline{y})$ and subtracting the results.

Doing so we have $\Delta_{uc}^{x,y} V_{n+1}^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc}) =$

$$\frac{1}{\psi_k}\Big[\beta\big[\sum_{ij \in U\times C} [\lambda_{ij}\Delta_{uc}^{x,y} V_n^{k,\pi}(\underline{x}, \underline{y} + \underline{e}_{uc}) + (y_{ij}^+ \wedge k)\,\eta\,\Delta_{uc}^{x,y} V_n^{k,\pi}(\underline{x} + \underline{e}_{ij} - \underline{e}_{uc}, \underline{y} - \underline{e}_{ij} + \underline{e}_{uc})]$$

$$- \mathbb{1}_{\{y_{uc}<k\}}\,\eta\, V_n^{k,\pi}(\underline{x}, \underline{y})$$

$$- \hat{\mu}_{uc}\,[p_{uc}\Delta_{uc}^{x,y}\Delta_{uc}^{y} V_n^{k,\pi}(\underline{x} - 2\underline{e}_{uc}, \underline{y} + \underline{e}_{uc}) + \Delta_{uc}^{x,y}\Delta_{uc}^{x,y} V_n^{k,\pi}(\underline{x} - 2\underline{e}_{uc}, \underline{y} + \underline{e}_{ij} + \underline{e}_{uc})]$$

$$+ \big(\psi_k - \sum_{ij \in U\times C} [\lambda_{ij} + (y_{ij}^+ \wedge k)\,\eta]\big)\,\Delta_{uc}^{x,y} V_n^{k}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})$$

$$+ \mathbb{1}_{\{y_{uc}<k\}}\,\eta\, V_n^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})\big]\Big], \tag{3.17}$$

where $y_{ij}^+ = y_{ij}$ for all $ij \neq uc \in U \times C$, and $y_{uc}^+ = y_{uc} + 1$. In a similar way, and by using action $a = sl$ (policy $\hat{\pi}$) in (3.14) quantities $\Delta_{sl}^{y} V_{n+1}^{k,\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y})$ and $\Delta_{sl}^{x,y} V_{n+1}^{k,\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y} + \underline{e}_{sl})$ can be computed. Next, to check property (3.15) for $n+1$, multiply (3.16) by $p_{uc}\,\hat{\mu}_{uc}$, and (3.17) by $\hat{\mu}_{uc}$ and add up the results. Similarly, multiply $\Delta_{sl}^{y} V_{n+1}^{k,\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y})$ and $\Delta_{sl}^{x,y} V_{n+1}^{k,\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y} + \underline{e}_{sl})$ by $p_{sl}\,\hat{\mu}_{sl}$ and $\hat{\mu}_{sl}$, respectively, and add up the results. Next, using the induction hypothesis and that $p_{uc}\,\hat{\theta}_{uc}\,\hat{\mu}_{uc} \geq p_{sl}\,\hat{\theta}_{sl}\,\hat{\mu}_{sl}$, after algebraic simplification it follows that

$$\hat{\mu}_{uc}\,[p_{uc}\Delta_{uc}^{y} V_{n+1}^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y}) + \Delta_{uc}^{x,y} V_{n+1}^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})]$$

$$- \hat{\mu}_{sl}\,[p_{sl}\Delta_{sl}^{y} V_{n+1}^{k,\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y}) + \Delta_{sl}^{x,y} V_{n+1}^{k,\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y} + \underline{e}_{sl})] \geq 0, \tag{3.18}$$

which establishes property (3.15) for $n+1$ for the case where $x_{uc}, x_{sl} \geq 2$. In a similar way, this property can be established for other cases (i.e., the remaining partition of the state space). Hence, a non-idling strict priority rule is optimal for all n. Next, taking the limit as $n \to \infty$ it follows that the finite horizon problem converges to the infinite horizon one both in policy and cost (see [126] Proposition 4.3.1). Furthermore, the convergence of the policy of the infinite-horizon discounted cost problem to that of average cost can easily be established (see [126] Corollary 7.5.10). Therefore, the underlying non-idling strict priority policy is optimal under the average cost setting

indexed by k (i.e., with $\cdot/M/k$'s in place of the $\cdot/M/\infty$) for any finite k. Since the result is true for any k, a convergence argument can be used to show that the result holds for the original problem with $k = \infty$. Notice that the existence of an optimal stationary policy for the original CTMS (i.e., when $k = \infty$) follows from the results of [56]. $\square$

CHAPTER 4

Dynamic Control in the W Service Network and Beyond

4.1 Introduction

The use of cross-trained workers (or flexible machines) in manufacturing or service sectors provides flexibility by dynamically shifting workers (workloads) to respond to volatile demands, machine/worker availabilities, congestion, etc. Typically, agents/workers are *partially flexible*, in that they are trained to serve a limited number of different requests (task types) so as to achieve a cost-effective level of flexibility.

The literature on the modeling and analysis of flexibility includes the following three themes: (1) *System Design* of specific paradigms for creating flexibility to maximize an objective, (2) *Server Scheduling and Control* policies to reap the benefits of flexibility, and (3) *Performance Analysis* of specific systems and policies. Our work contributes to all three themes, especially the second theme.

System design, the first theme, motivates the development of methodology to determine which capabilities a server should be endowed with (see, for instance, [84], [2], [71], [72], [81], [80], [15], [26], and [7]). We first analyze parallel (Markovian) queueing systems with general structures (i.e., arbitrary number of queues and servers in parallel with general skill/capability sets for the servers) to prove properties such

as stability. Then we focus on the "W" paradigm/structure for parallel operations that are "make-to-order."

To motivate the "W" paradigm, consider the small customer support center illustrated in Fig. 4.1. In this example, both agents can handle phone calls. However, only one of them is responsible for supporting customers through the internet using chat and email. The other agent is provided the resources to handle postal mail and faxes. We refer to the queueing structure of Fig. 4.1 as a "W" queueing network (since it forms a "W" with respect to the server skills and workflow). For the manager of the system illustrated in Fig. 4.1, different request types have different response time urgencies. For instance, a quick reply to a chat or an email request is often more important than a fast response to a postal mail or to a fax. Phone calls on hold (waiting in queue) are also usually more urgent than a mail or a fax request.

Generally, in such systems, a per unit of time cost (or relative weight) of h'_i can be assigned to holding a request of type i. Additionally, the servers are usually *heterogenous*: they have different skill levels (service rates) in serving different job types. In general, one can model the service of a request of type i by server j as occurring with a rate of $\mu'_{ji} \geq 0$ (where zero indicates that server j lacks skill i). Moreover, servers might be subject to *stochastic disruptions* occurring with a rate of $\theta'_j \geq 0$ for server j, which represents the time lost due to an IT disruption, unplanned absences (e.g. unexpected meetings), etc. When disrupted, server j returns to a working state after an expected r'^{-1}_j units of time, which represents its average "repair" time. Assuming a type i request comes to the system at rate λ'_i, the manager of such a system needs to know how to assign the agents to different requests in *real-time* to obtain good performance and extract the most benefit from the partial flexibility of the servers.

Conceptually, the "W" structure can be observed in many systems in practice. One of the situations where a "W" structure may naturally arise is where tasks performed by the servers have a wide variety and can be classified as tasks that are server specific and tasks that are shared between servers. Consider, for example, a

small clinic with a physician and a nurse working together. There is a set of tasks that would be performed by the nurse (e.g. taking blood pressure and other diagnostic tests, administering medications, and basic treatments) and there is also a separate set of tasks that would be performed by the physician (e.g. diagnosis of diseases and injuries, prescribing medications and treatments, and performing higher skill medical procedures). Additionally, there is a set of tasks that could be performed by either the nurse or the physician, depending on the workloads of the nurse and the physician (e.g., diagnostic tests, bandaging, and giving home self-care or follow-up instructions).

The "W" structure may also arise from considerations of demand workload and service capacity. For example, the shared demand type may represent the demand class for which a server cannot provide enough capacity, and thus, capacity can be shifted via cross-training. Moreover, there are often tasks that are not cost effective to cross-train. This may be caused by the training/certification expense of the skill, the difficulty in obtaining workers competent at that skill, or the infrastructure and layout that makes the cross-functionality ineffective.

We contribute to the first theme of flexibility research, system design, by showing that the "W" structure achieves most of the potential performance with two servers and three job types, supporting the notion that *a little flexibility goes a long way* (which has been a theme of several papers such as [84] and [15] for some different structures). Considering the expense of cross-training servers and, more importantly, application-specific obstacles to cross-train certain task types, the frugality of the "W" design makes it widely useful in application.

We contribute to the second theme of flexibility research, control, by generating insights into effective mechanisms for the control of servers in the "W" design as well as systems with any general structure. Specifically, for the "W" structure, we rigorously establish a partial characterization of the $c\mu$ rule (i.e., the weighted shortest processing time policy) as an optimal policy under certain operating conditions. We also develop a high-performance heuristic index policy, termed as *"Largest Expected*

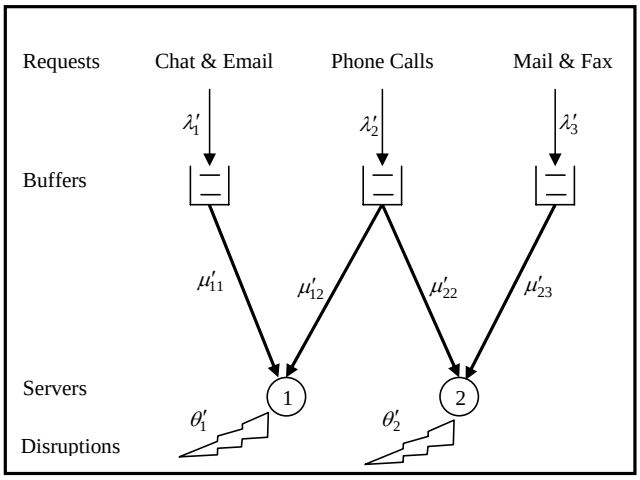


Figure 4.1: An example of a small customer support center (the "W" structure).

Waiting Cost" (LEWC), and benchmark it relative to the optimal policy for a large test suite. The proposed LEWC index policy, however, is not specific to the "W" design and can be implemented in any parallel queueing system.

Even after ignoring possible disruptions, the control problem that we consider in this chapter is a difficult and still an open area of research. For instance, [16] considered the "N" structure (with reliable servers), a special case of the "W" with the third demand stream removed, and noted that even for the "N" *"the problem of finding a control policy that minimizes a cost associated with holding jobs in the system is notoriously difficult.*" The "W" model is a significant departure from the "N," because it has two partially flexible servers, whereas the "N" has two extremes: one inflexible and one fully flexible server. Server disruptions further complicate the problem. In addition to identifying sufficient conditions under which the well-known greedy $c\mu$ policy is optimal, our numerical analysis provides further insights for situations where those conditions do not hold. Particularly, the optimal policy is a *state-dependent threshold type policy* characterized by four switching surfaces in the cases studied.

Addressing the system design agenda of the first theme, [81], [78], and [80] have developed methodologies such as Structural Flexibility and Capability Flexibility for

estimating the better of alternative cross-training architectures with respect to mean waiting time. To test these methods, the above papers primarily used the Longest Queue (LQ) as the control policy. In this chapter, we propose LEWC as a more effective policy. It should be noted that even for a particular structure such as "W," performance analysis (the third theme) under the optimal policy is difficult. Thus, we provide a careful MDP-based numerical benchmarking study that gives insights into the optimal policy as well as LEWC, LQ, $c\mu$, and generalized $c\mu$ ($Gc\mu$) with quadratic holding cost (also referred as Max-Weight). We find that not only does the LEWC heuristic clearly outperforms LQ, $c\mu$, and $Gc\mu$, but it is also a near optimal policy with a relatively small optimality gap. Moreover, we establish its stability. Since LEWC can be used for the control of servers in systems with any flexibility structure, the obtained results introduce LEWC as a promising policy for future research into the design of flexible structures with arbitrary topologies. This is particularly useful because the comparison of alternative flexibility/queueing designs under their optimal control policies is computationally intractable for large systems.

The rest of this chapter is organized as follows. Section 4.2 briefly reviews some related studies. Section 4.3 formulates the problem using an MDP framework and identifies some attributes applicable to a parallel queueing system with a general flexibility structure. Section 4.4 presents the results on the "W" structure, describes the proposed LEWC heuristic, and extensively tests its performance.

4.2 Literature Survey

When there is a single server in the system that is fully flexible and has memoryless service times, [22] and [156] show that the well-known $c\mu$ policy is optimal. The $c\mu$ rule is a very intuitive and easy control policy to implement; however, it may perform poorly when partial flexibility is introduced, as is the case with the "W." It remains, however, optimal under some conditions. For instance, [39] prove the optimality of

the $c\mu$ rule for an "N" structure under some special conditions. [155] shows the optimality of $c\mu$ for systems without disruptions where servers collaborate on jobs and special conditions are satisfied.

In parallel systems, which is our focus, the literature mainly considers the control problem in the heavy-traffic regime (see, for instance, [151], [61], [62], [16], [21], [105], [102], and [17]). The literature, however, lacks policies that are effective for a *wide range of utilizations*. Our target in this chapter is on systems in the utilization range of 70% to 90%.

The problem of dynamically assigning servers to jobs has also been studied under the throughput maximization objective (see, for instance, [8], [9], [12], [33], and [6]). Among these papers, [6] is most related to our work since it also allows for disruptions. However, throughput maximization is appropriate only for systems in which delay is not a major concern, and in most cases it is an easier problem to analyze.

Work on the benefit of flexibility to compensate for the risk of disruptions is also related to our work. For this stream of research, we refer interested readers to [6], [122], [123], and the references therein.

4.3 General Characteristics

This section addresses general Markovian parallel queueing structures with partially flexible and possibly unreliable servers. That is, we consider Markovian parallel queueing systems with an arbitrary number of servers, arbitrary number of customer classes, and arbitrary flexibility structures. We allow even more generality by allowing for a stochastic disruption/repair process unique to each server. We first describe our model, and then formulate it using an MDP framework.

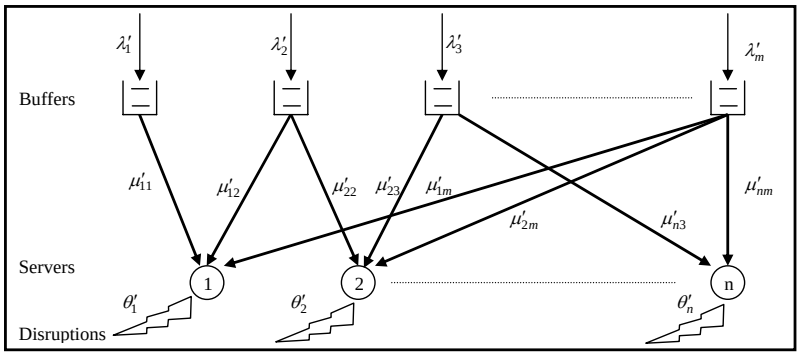


Figure 4.2: A general parallel queueing system with server disruptions and arbitrary flexibility structure.

4.3.1 The Model

Consider a queueing system represented by a bipartite graph $G = (\mathcal{N}, \mathcal{E})$ where $\mathcal{N}$ is partitioned to two finite sets: $\mathcal{N}_c = \{1, \cdots, m\}$ for customer/jobs classes, and $\mathcal{N}_s = \{m+1, \cdots, m+n\}$ for servers/machines (see Fig. 4.2). (A labeling $\{1, \cdots, n\}$ might be used for servers when it does not generate a confusion.) Arrivals of customers of class $i \in \mathcal{N}_c$ follow a Poisson process with rate $\lambda'_i \in \mathbb{R}^+$, and server $j \in \mathcal{N}_s$ can serve a customer of class $i \in \mathcal{N}_c$ with an exponentially distributed amount of time with rate $\mu'_{ji} \in \mathbb{R}^+$. In the graph G, $(i,j) \in \mathcal{E} \subseteq \mathcal{N}_c \times \mathcal{N}_s$ if, and only if, $\mu'_{ji} > 0$. We let $\mathcal{S}_j = \{i : (i,j) \in \mathcal{E}\}$ denote the skill set or "capabilities" of server j and $\mathcal{S}_i^{-1} = \{j : (i,j) \in \mathcal{E}\}$ denote the servers capable of serving class i. To allow for server unreliability, disruptions to server $j \in \mathcal{N}_s$ occur according to a Poisson process with rate $\theta'_j \geq 0$ (equality holds if server j is completely reliable). Note that we focus on systems for which disruptions occur at the same rate whether or not the server is in use. For example, unplanned employee absence, a power outage, or an economic disruption may happen independently of server idleness. Once a server is disrupted, it immediately undergoes a repair process that takes an exponentially distributed amount of time with rate $r'_j > \theta'_j$ for server j. All above-mentioned stochastic processes are considered to be independent of each other.

Let $\mathbf{h}' = (h'_1, \cdots, h'_m)$, where h'_i denotes the per unit time (inventory) holding

cost associated with holding a customer of class i. The objective is to find an optimal resource allocation (or server assignment) policy to minimize the average holding cost of the system assuming that the servers cannot collaborate on the same job (unless otherwise mentioned), but service preemption is permitted. To achieve this goal, let $\mathbf{X}^\pi(t) = (X_1^\pi(t), \cdots, X_m^\pi(t))$ where $X_i^\pi(t)$ denotes the number of class i customers in the system at time t under policy π. A policy is then optimal if it achieves the following optimal cost:

$$Z^* = \inf_{\pi \in \Pi} Z^\pi = \inf_{\pi \in \Pi} \Big\{ \sum_{i \in \mathcal{N}_c} h_i' L_i^\pi \Big\}, \tag{4.1}$$

where Π is the set of all admissible policies, and L_i^π denotes the long-run average number of class i customers in the system under policy π. This latter measure can be computed as:

$$L_i^\pi = \limsup_{T \to \infty} \frac{1}{T} \int_0^T E[X_i^\pi(s)]\, ds. \tag{4.2}$$

4.3.2 Formulation of the Markov Decision Process

For $j \in \mathcal{N}_s$, let $a_j(t) = 1$ denote that server j is available (i.e., not disrupted) at time t and let $a_j(t) = 0$ otherwise. The state of the system is then a vector $\tilde{\mathbf{X}}(t) = (\mathbf{X}(t), \mathbf{a}(t))$ with state space $S = \mathbb{Z}^{+m} \times \{0,1\}^n$, where $\tilde{\mathbf{X}}_i(t) = X_i(t) \in \mathbb{Z}^+$ for $i \in \mathcal{N}_c$ and $\tilde{\mathbf{X}}_j(t) = a_j(t) \in \{0,1\}$ for $j \in \mathcal{N}_s$. We use uniformization (see [99]) to formulate the discrete time equivalent of the problem. Since $\theta_j' < r_j'$, we use the uniformization factor $\psi = \sum_{i \in \mathcal{N}_c} \lambda_i' + \sum_{j \in \mathcal{N}_s} r_j' + \sum_{j \in \mathcal{N}_s} \max_{i \in \mathcal{N}_c} \{\mu_{ji}'\}$ (where $0 < \psi < \infty$). Let $\lambda_i = \lambda_i'/\psi$, $\mu_{ji} = \mu_{ji}'/\psi$, $\theta_j = \theta_j'/\psi$, and $r_j = r_j'/\psi$ denote the parameters after uniformization corresponding to the transition probabilities in the underlying discrete Markov chain. Also, let α be a continuous time discount rate and ξ be an exponential random variable with rate ψ denoting the length of one unit of time in the corresponding discrete Markov chain. The equivalent discount factor in

discrete time is then:

$$\beta = E[e^{-\alpha\,\xi}] = \int_0^\infty (e^{-\alpha t})\,(\psi e^{-\psi t})dt = \frac{\psi}{\alpha+\psi} \tag{4.3}$$

Also, since the state of the system does not change in one period of the discrete time version, the equivalent instantaneous one period cost is:

$$\mathbf{h}\mathbf{X}^T = E\Big[\int_0^\xi \mathbf{h}'\mathbf{X}^T e^{-\alpha t}\,dt\Big] = \frac{1-\beta}{\alpha}\,\mathbf{h}'\mathbf{X}^T = \frac{\mathbf{h}'}{\psi+\alpha}\mathbf{X}^T, \tag{4.4}$$

and so $\mathbf{h} = \mathbf{h}'/(\psi+\alpha)$. The finite-horizon optimal expected discounted cost can then be computed using the following optimality equation defined for every $\tilde{\mathbf{X}} \in S$ and $n \in \mathbb{Z}^+$:

$$\begin{aligned}
&V_{n+1,\beta}(\tilde{\mathbf{X}}) = \\
&\mathbf{h}\mathbf{X}^T + \beta\left[\sum_{i\in\mathcal{N}_c}\lambda_i V_{n,\beta}(A_i\tilde{\mathbf{X}}) + \sum_{j\in\mathcal{N}_s}\left[\theta_j\,a_j\,V_{n,\beta}(B_j\tilde{\mathbf{X}}) + r_j\,(1-a_j)\,V_{n,\beta}(R_j\tilde{\mathbf{X}})\right]\right. \\
&+\min_{\mathbf{u}\in\mathcal{U}(\tilde{\mathbf{X}})}\left\{\sum_{i\in\mathcal{N}_c}\sum_{j\in\mathcal{N}_s}\mathbb{1}\{u_j=i\}\,\mu_{ji}V_{n,\beta}(D_i\tilde{\mathbf{X}})\right. \\
&\left.\left.\quad+\left(1-\sum_{i\in\mathcal{N}_c}\lambda_i-\sum_{i\in\mathcal{N}_c}\sum_{j\in\mathcal{N}_s}\mathbb{1}\{u_j=i\}\,\mu_{ji}-\sum_{j\in\mathcal{N}_s}\left[\theta_j\,a_j+r_j\,(1-a_j)\right]\right)V_{n,\beta}(\tilde{\mathbf{X}})\right\}\right],
\end{aligned} \tag{4.5}$$

where $V_{n,\beta}(\tilde{\mathbf{X}})$ represents the optimal cost of an n-period problem starting at state $\tilde{\mathbf{X}}$, $\mathbb{1}\{\cdot\}$ is the indicator function, and the initial condition is $V_{0,\beta}(\tilde{\mathbf{X}}) = 0$ for every $\tilde{\mathbf{X}} \in S$. In this optimality equation, the arrival, departure, repair, and breakdown state transition operators for $i \in \mathcal{N}_c$ and $j \in \mathcal{N}_s$ are denoted by $A_i\tilde{\mathbf{X}} = \tilde{\mathbf{X}} + \mathbf{e}_i$, $D_i\tilde{\mathbf{X}} = \tilde{\mathbf{X}} - \mathbf{e}_i$, $R_j\tilde{\mathbf{X}} = \tilde{\mathbf{X}} + \mathbf{e}_j$, and $B_j\tilde{\mathbf{X}} = \tilde{\mathbf{X}} - \mathbf{e}_j$, respectively, where $\mathbf{e}_i$ ($\mathbf{e}_j$) is a vector with the same dimension as S with a one in ith (jth) position and zeros elsewhere. Moreover, the control action is the vector $\mathbf{u} = (u_j \in \mathcal{N}_c \cup \{0\}, \forall j \in \mathcal{N}_s)$ where $u_j = i \in \mathcal{N}_c$ if server j is assigned to serve class i, and $u_j = 0$ if it is not assigned to any class. The set of admissible control actions at state $\tilde{\mathbf{X}}$ is denoted by

a set of vectors $\mathcal{U}(\tilde{\mathbf{X}})$, where: $\mathcal{U}(\tilde{\mathbf{X}}) =$

$$\left\{\mathbf{u} = (u_j \in \mathcal{N}_c \cup \{0\} \text{ s.t. } \forall i \in \mathcal{N}_c : \mathbb{1}\{u_j = i\} \leq a_j \, \mathbb{1}\{i \in \mathcal{S}_j\}\,, \sum_{j \in \mathcal{N}_S} \mathbb{1}\{u_j = i\} \leq X_i)\right\}. \tag{4.6}$$

That is, server j cannot be assigned to class i if it is disrupted, if it lacks skill i, or if the number of class i jobs is insufficient.

Similar to (4.5), the optimal average inventory holding cost can be computed using an MDP with the following average-cost optimality equation:

$$\begin{aligned} J(\tilde{\mathbf{X}}) + Z_U^* = \frac{\mathbf{h}'}{\psi}\,\mathbf{X}^T + \sum_{i \in \mathcal{N}_c} \lambda_i J(A_i \tilde{\mathbf{X}}) + \sum_{j \in \mathcal{N}_s} \left[\theta_j \, a_j \, J(B_j \tilde{\mathbf{X}}) + r_j \, (1 - a_j) \, J(R_j \tilde{\mathbf{X}})\right] \\ + \min_{\mathbf{u} \in \mathcal{U}(\tilde{\mathbf{X}})} \left\{ \sum_{i \in \mathcal{N}_c} \sum_{j \in \mathcal{N}_s} \mathbb{1}\{u_j = i\} \, \mu_{ji} J(D_i \tilde{\mathbf{X}}) \right. \\ \left. + \left(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{i \in \mathcal{N}_c} \sum_{j \in \mathcal{N}_s} \mathbb{1}\{u_j = i\} \, \mu_{ji} - \sum_{j \in \mathcal{N}_s} \left[\theta_j \, a_j + r_j \, (1 - a_j)\right]\right) J(\tilde{\mathbf{X}}) \right\}, \end{aligned} \tag{4.7}$$

where $J(\tilde{\mathbf{X}})$ is a relative cost function, Z_U^* denotes the optimal per period average cost in the uniformized problem, and $Z^* = \psi \, Z_U^*$ is the optimal per period average cost of the original problem. The next section first analyzes the stability conditions of the the general queueing system under consideration. Then it provides another method to compute the optimal average cost and the relative function $J(\tilde{\mathbf{X}})$ using the finite-horizon version of the problem (i.e., value iteration).

4.3.3 Stability

It is important first to identify the stability region of the system for several reasons. In addition to several interesting theoretical considerations, it provides an important practical design guideline. We define the general queueing network under consideration to be stabilizable if, and only if, there exists a policy $\pi \in \Pi$ such that $Z^\pi = \sum_{i \in \mathcal{N}_c} h_i' \, L_i^\pi < \infty$. This is equivalent to the existence of a *finite* mean

equilibrium distribution of the underlying stochastic process $\{\mathbf{X}(t), t \geq 0\}$. To check the stability of the underlying system with partially flexible and unreliable servers, we develop and implement the following Linear Program (LP) (in the spirit of [62] and [6]). Our LP maximizes the minimum *"excess service capacity,"* τ, that can be provided for *all* customer classes.

LP 1:

$$\text{Max} \quad \tau \tag{4.8}$$

Subject to:

$$\sum_{j \in \mathcal{S}_i^{-1}} y_{ji} \left(\frac{r_j}{\theta_j + r_j}\right) \mu_{ji} \geq \lambda_i + \tau \qquad \forall i \in \mathcal{N}_c, \tag{4.9}$$

$$\sum_{i \in \mathcal{S}_j} y_{ji} \leq 1 \qquad \forall_j \in \mathcal{N}_s, \tag{4.10}$$

$$y_{ji} \geq 0 \qquad \forall j \in \mathcal{N}_s, \forall i \in \mathcal{S}_j. \tag{4.11}$$

In this LP, we introduce the decision variable y_{ji} ($j \in \mathcal{N}_s, i \in \mathcal{S}_j$) to denote the long-run proportion of time that server j is "assigned" to work on class i (including the times during which server j is disrupted) when the arrival rate of class i is $\lambda_i + \tau$. Notice that (using either Renewal Theory or a two-state Markov chain model of disruption and repair process) server j in steady state is available $\frac{r_j}{\theta_j + r_j}$ percent of the time. Thus, $y_{ji}\left(\frac{r_j}{\theta_j + r_j}\right)$ represents the long-run proportion of the time that server j is available and working on class i (when the arrival rate of class i is $\lambda_i + \tau$); and $y_{ji}\left(\frac{r_j}{\theta_j + r_j}\right)\mu_{ji}$ is the corresponding long-run average capacity offered to class i by server j given y_{ji}. Hence, from constraint (4.9) we see that objective function (4.8) maximizes the minimum excess capacity among all classes. Constraint (4.10) (together with (4.11)) sets an upper bound for the total fraction of time that a server can be assigned to a specific class. The following theorem, based on fluid model analysis (see, for instance, [34]) and similar to some results presented in the literature (see for instance, [6]), relates the above LP to the stabilizability of the system. This theorem provides a tool to ensure that the class of finite cost policies is not empty,

and hence the optimization in (4.1) is of interest. See Online Appendix A for all of the proofs.

Theorem 2 (Stability) *Let τ^* be the optimal objective value of LP 1. Then:*
(i) The system is stabilizable (i.e., $\exists \pi \in \Pi$ s.t. $Z^\pi < \infty$) if $\tau^* > 0$.
(ii) The system is not stabilizable (i.e., $\forall \pi \in \Pi : Z^\pi = \infty$) if $\tau^* < 0$.

Now that we have a tool to check stabilizability, we can take one step further and (1) guarantee the existence of an optimal *stationary* policy, and (2) establish the convergence of the finite-horizon problem to the average-cost case (both in the cost and in the policy). Indeed, we can establish a convenient alternative approach to find an optimal average-cost policy by stating that (1) it is sufficient to restrict attention to the class of stationary policies, and (2) solving the finite-horizon version of the problem defined in (4.5) can provide both the average-cost optimal value Z^* and the average-cost optimal policy π^*.

Theorem 3 (Stationary Policy & Convergence) *If $\tau^* > 0$, then:*

(i) There exists an average-cost optimal stationary policy.

(ii) The optimal average cost can be computed by:
$Z^* = \inf_{\pi \in \Pi} \{\sum_{i \in \mathcal{N}_c} h_i' L_i^\pi\} = \lim_{\beta \to 1^-} \lim_{n \to \infty} \psi\,(1-\beta) V_{n,\beta}(\tilde{\mathbf{X}})$.

(iii) The relative cost function $J(\tilde{\mathbf{X}})$ defined in (4.7) satisfies:
$J(\tilde{\mathbf{X}}) = \lim_{\beta \to 1^-} \lim_{n \to \infty} [V_{n,\beta}(\tilde{\mathbf{X}}) - V_{n,\beta}(\mathbf{0})]$.

(iv) Let $\pi_{n,\beta}$ denote an optimal policy for the n-period (discounted cost) problem. Then any limit point π_β of the sequence $\{\pi_{n,\beta}\}_{n \geq 1}$ (as $n \to \infty$) is optimal for the infinite-horizon discounted cost. Moreover, any limit point of the sequence $\{\pi_\beta\}_{\beta \in (0,1)}$ (as $\beta \to 1^-$) is average-cost optimal.

In the search for effective mechanisms to control the servers, we are able to restrict our attention to the class of policies that do not allow for unforced idling. This is shown in Appendix A (see Lemma 4 and Proposition 7), where we establish this result based on a proof of the monotonicity of the value function.

4.4 The "W" Structure

In the previous section, we presented some characteristics applicable to any Markovian parallel queueing system with an arbitrary server flexibility structure. In this section, to develop more insights, we consider a special structure forming a "W" (see Fig. 4.1 or Structure 4 in Fig. 4.3). This structure is an especially effective paradigm for systems with three demand types and two servers. It should be noted that the "N" structure, widely studied in the literature, is a special case of a "W" with $\mu'_{23} = \lambda'_3 = 0$.

The next section shows that the "W" is an efficient design that requires only a little cross-training to achieve a performance almost as good as any design with two servers and three job types. Since cross-training the servers is costly (and sometimes infeasible) in practice, this observation shows that; for systems with three demand types and two servers, instead of fully cross-training every server, it is sufficient to make them capable to serve a shared task in addition to their dedicated/fixed one and form a "W" structure.

4.4.1 The "W" Structure: An Efficient System Design

From a system design perspective, it is crucial to understand the effective ways of cross-training servers. Please note that in this section we focus on congestion (and mean wait), so all holding costs are set to one. To understand the design problem, consider the various possible designs with three customer classes and two servers illustrated in Fig. 4.3. These six structures progressively add skills, except Structures 2, 3, and 4 (the "W"), which have the same number of skills. Thus, Structures 2, 3, and 4 also allow us to explore the sensitivity with respect to *where* the fourth skill is added. In Structures 3 and 4 the class with the highest arrival rate is the shared one, but it is not the case in Structure 2. Structure 2 is indeed a "W," where the shared task is not the one with the highest arrival rate (i.e., the middle class). The goal is to find an *efficient design* among these five structures. In other words, to improve

the design of Structure 1, we address two questions: (1) *where* should one implement flexibility/cross-training?, and (2) *how many* additional skills are adequate to get a reasonably good performance?

To answer these questions, we compare the performance of the above-mentioned structures under their optimal policies in various test suites (parameter settings) as presented in Table 4.3 (see Online Appendix C). Notice that, considering the built-in symmetry in our test suites (symmetry between classes 1 and 3 as well as the symmetry in the speed of a server in serving different classes), the six structures considered in Fig. 4.3 cover all the possible designs; any other (stabilizable) structure is homomorphic to one of these six structures. Fig. 4.4 summarizes our computational results by depicting the optimal long-run average number of customers in each of these six structure for our test suite and under various congestion factors (ρ in Table 4.3). The mean (i.e., long-run average) number of customers (or jobs) in the system under the optimal policy is computed by numerically solving the average-cost MDP optimality equation (4.7) with $h'_i = 1\,(\forall i \in \mathcal{N}_c)$.

The results depicted in Fig. 4.4, which is a summary of optimally solving 6 (structures)$\times$ 4 (suites) $\times$ 9 (congestion factors) = 306 problem instances, confirm that (a) flexibility usually has a diminishing rate of return, (b) a little flexibility can go a long way and (c) it usually matters where to add the additional flexibility, which has been elaborated on in studies such as [84], [71], [81], and [15]. The primary intent of this section is, however, to reveal the following insight about the "W."

- **Insight.** Structure 4, the "W" (or to be precise the "W" with the proper task being shared) is an efficient design where a little cross-training can achieve most of the flexibility of a fully flexible network (i.e., Structure 6). In test suites 2, 3, and 4, the "W" is almost as good as Structure 6 and in test suite 1, the "W" is still an efficient architecture. This observation is especially important considering the expense of cross training servers in most practical situations and reveals the benefit of implementing a "W" structure.

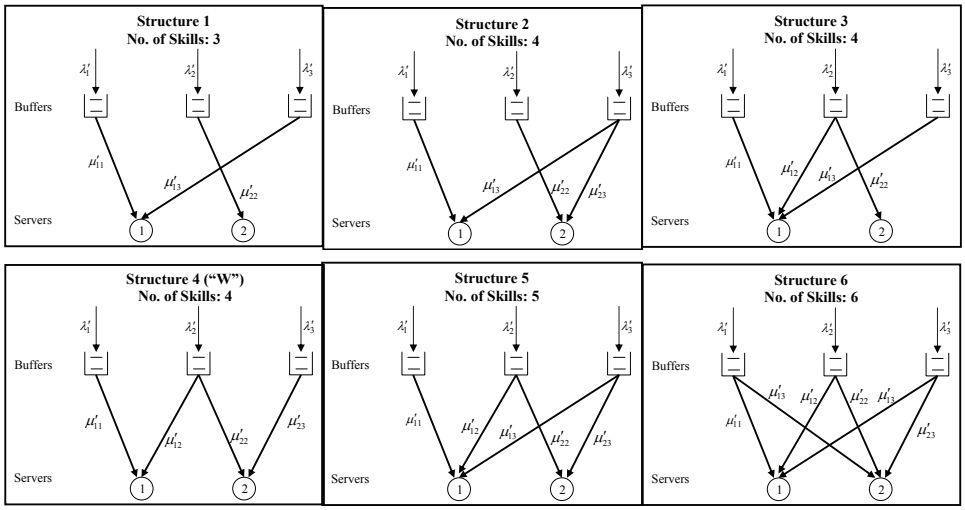


Figure 4.3: Various possible structures with $|\mathcal{N}_c| = 3$ and $|\mathcal{N}_s| = 2$.

It should be noted that similar characteristics have been shown in the literature for chaining (see for instance [71] and [84]) and tailored pairing ([15]), but the "W" is not a special case of those structures. Concurrent research in [7] takes an alternate approach to analyzing the "W" and other structures with respect to throughput.

4.4.2 Dynamic Control of Servers in the "W" Structure

The previous section examined the benefit of implementing a "W" structure; however, this benefit cannot be fully achieved without efficiently assigning servers to jobs in real-time. Hence, the remaining question is: *what control policy should be used in real-time to extract the most benefit from the limited flexibility of servers in this design?* The answer to this question will also provide insight into the control of more complex queueing structures with partially flexible servers. We first state a corollary of Theorem 2 to partially characterize the stability region of a "W" design in more insightful expressions.

Corollary 1 (Stability of "W") *Consider a "W" structure under stochastic disruptions (or without them as a degenerate case). Let* $\rho_1 = \lambda_1/(\mu_{11}\frac{r_1}{r_1+\theta_1})$ *and* $\rho_3 =$

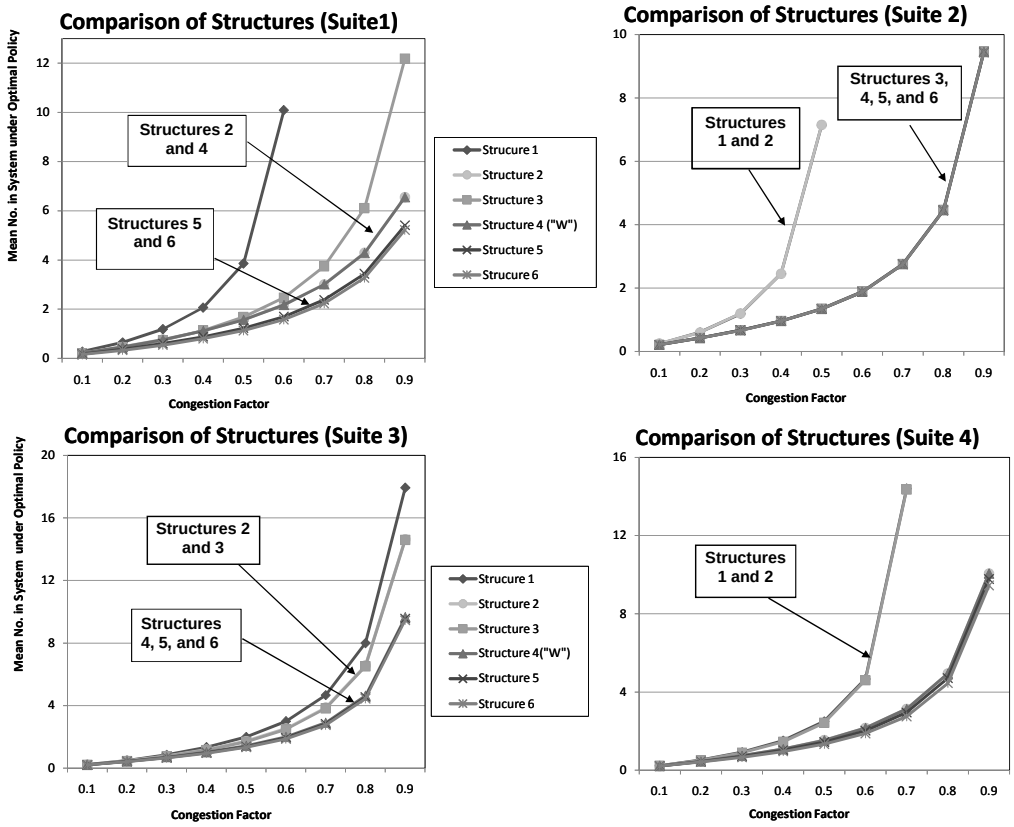


Figure 4.4: Comparison of possible structures under four suites of parameters using the optimal policies.

$\lambda_3/(\mu_{23}\frac{r_2}{r_2+\theta_2})$. *Also, define effective service rates of the shared task as* $\mu_{12}^{eff} = \mu_{12}\frac{r_1}{r_1+\theta_1}$ *and* $\mu_{22}^{eff} = \mu_{22}\frac{r_2}{r_2+\theta_2}$. *The system is not stabilizable if* $\max\{\rho_1, \rho_3\} > 1$. *On the other hand, if* $\max\{\rho_1, \rho_3\} < 1$, *the system is stabilizable if* $(1-\rho_1)\,\mu_{12}^{eff} + (1-\rho_3)\,\mu_{22}^{eff} > \lambda_2$.

Now we characterize the optimal control policy. Hereafter, we assume the system under consideration is stabilizable. The following theorem shows the optimality of prioritizing the *fixed task before the shared* for every server under certain conditions. This policy is an analogous to the well-known $c\mu$ ($h\mu$ in our notation) rule as a strict priority ordering for every server.

Theorem 4 (Optimality of the $c\mu$ Strict Priority: Fixed before Shared) *For a "W" structure with stochastic disruptions (or without them as a degenerate case), if* $h_1'\,\mu_{11}' \geq h_2'\,\mu_{12}'$, $h_3'\,\mu_{23}' \geq h_2'\,\mu_{22}'$, *and either (i) server collaboration is allowed, or*

(ii) server collaboration is disallowed but $\mu'_{12} \geq \mu'_{11}$ *and* $\mu'_{22} \geq \mu'_{23}$ *hold, then the* $c\mu$ *priority rule is optimal for each server. That is, there exists an optimal policy under which every server, when not disrupted and regardless of the other server's allocation or disruption state, prioritizes its fixed task before the shared task whenever its fixed queue is not empty.*

The *"Fixed before Shared"* policy described in the Theorem 3 can be viewed as an extension of the $c\mu$ rule for systems with partially flexible and unreliable servers. Indeed, the above theorem shows that this extension of the $c\mu$ policy is optimal for the "W" when the $c\mu$ index ($h\mu$ in our terminology) gives priority to the fixed task for each server (even when servers are unreliable). Under the conditions specified in Theorem 4, using the $c\mu$ strict priority rule for a server cannot bring the ill side effect of underutilizing the other server because of the specific flexibility structure. In other words, under these conditions, the $c\mu$ policy is *starvation free*; it maximizes the amount of job available to the other server, and hence, remains optimal. This insights might also hold for larger systems where $c\mu$ priorities are toward the fixed tasks for all servers. One nice feature of the above policy (i.e., Fixed before Shared) is that it defines a prescriptive rule for each server regardless of the other server's allocation or disruption state. This feature removes the need for servers to communicate in real-time and provides a static (i.e., state-independent) rule that is easy to implement.

Our extensive MDP-based numerical computations show that the optimal policy is complex in general when $c\mu$ priorities are not toward the fixed tasks. Relaxing all such assumptions, we observe from our extensive numerical examples that the optimal policy for a general "W" structure with server disruptions is a *state-dependent threshold type policy* which can be defined by four switching surfaces. See Online Appendix B for a detailed discussion on this observation and for numerical examples supporting it.

4.4.3 An Efficient Heuristic Policy: Largest Expected Workload Cost (LEWC)

In practice, to be implementable, a policy must be easy enough to use. In the experience of the authors, managers and researchers working with on-demand service centers usually believe that a simple policy such as LQ is preferable to the cost/effort of implementing a complex policy in real-time (see [72] and [81]). However, our investigation has revealed that the popular LQ policy does not perform well in many situations. Moreover, as the previous section revealed, the optimal policy is complex and hard to be implemented in real-time in practice. Therefore, in this section, we develop a heuristic policy that is both easy-to-implement and highly effective.

This policy balances the expected workload cost of queues. Indeed, this heuristic prescribes that every server (whenever not disrupted) in every decision epoch should prioritize serving the queue with the *Largest Expected Workload Cost (LEWC)*, regardless of the allocation or availability (i.e, disruption state) of other servers. Under this policy, a server does not need to know all the queue lengths. Rather, each server needs visibility only of her/his duty area (skill set) to decide which queue to serve. Moreover, this policy eliminates the need for communication between servers, since each server can perform her/his job without the knowledge about the other servers' allocations, availabilities, or workloads. As a result, a manager can prescribe a rule to each server *in advance* and ensure good overall performance. In large networks more general than the "W" this is a significant advantage. However, this policy is still dynamic and requires different actions for each server depending on the real-time length of the queues within the server's skill set.

To develop this policy, we first slightly modify LP 1 presented in Section 4.3.3; we call the new program LP 2. The objective of this LP (applicable to any general network and not only the "W") is to find allocations y_{ji} that maximize the minimum *percentage* excess capacity, $\tilde{\tau}$, among all queues.

LP 2:

$$\text{Max} \quad \tilde{\tau}$$

Subject to:

$$\sum_{j \in \mathcal{S}_i^{-1}} y_{ji} \left(\frac{r_j}{\theta_j + r_j}\right) \mu_{ji} \geq \lambda_i \left(1 + \tilde{\tau}\right) \qquad \forall i \in \mathcal{N}_c, \tag{4.12}$$

$$\sum_{i \in \mathcal{S}_j} y_{ji} \leq 1 \qquad \forall_j \in \mathcal{N}_s, \tag{4.13}$$

$$y_{ji} \geq 0 \qquad \forall j \in \mathcal{N}_s, \forall i \in \mathcal{S}_j. \tag{4.14}$$

Next, for each queue i (with queue length x_i), we develop an index $\mathcal{I}_i(x_i)$ to approximate the expected workload cost of that queue. We call this the *"LEWC index"* and define it as:

$$\mathcal{I}_i(x_i) = \frac{h_i \times x_i}{\sum_{j \in \mathcal{S}_i^{-1}} y_{ji}^* \left(\frac{r_j}{\theta_j + r_j}\right) \mu_{ji}} \ , \tag{4.15}$$

where y_{ji}^* are the solution to LP 2, and $\mathcal{S}_i^{-1}$ represents the set of servers able to serve queue i. In fact, if all servers that can work on queue i are assigned to work there based on the steady state allocations obtained from LP 2, a single job in the first position of queue i will take $[\sum_{j \in \mathcal{S}_i^{-1}} y_{ji}^* \left(\frac{r_j}{\theta_j + r_j}\right) \mu_{ji}]^{-1}$ units of time to be served (assuming work sharing is permitted). Since x_i jobs are in queue i, it will take approximately (ignoring the waiting times) $x_i \times [\sum_{j \in \mathcal{S}_i^{-1}} y_{ji}^* \left(\frac{r_j}{\theta_j + r_j}\right) \mu_{ji}]^{-1}$ units of time to serve all the jobs in queue i. This generates a workload cost of $\mathcal{I}_i(x_i)$ for queue i. It should be clear that the LEWC index also accounts for other system parameters, such as arrival rates, disruption rates, and repair rates through the optimal solutions y_{ji}^*. Therefore, LEWC incorporates not only the load balancing logic of LQ and the greedy cost minimization of $c\mu$, but also considers utilizations via solutions y_{ji}^*. The LEWC heuristic policy follows.

LEWC Algorithm:

- **Step 1:** Solve LP 2 to obtain the optimal allocations y_{ji}^*.
- **Step 2:** At the current state, $\tilde{\mathbf{X}}$, use (4.15) to compute indexes $\mathcal{I}_i(x_i)$ for

all queues (i.e., $i \in \mathcal{N}_c$). Then assign each available server j to the queue $i_j^* = \text{argmax}_{i \in \mathcal{S}_j} \mathcal{I}_i(x_i)$, i.e., to the queue with the largest LEWC index among the queues that it can serve. If two or more queues have the same index, break the tie by assigning the server to the queue with the smallest label (i.e., the left most queue in our diagrams).

The following theorem states that our proposed policy stabilizes the system, if the system is stabilizable (i.e., if there exists a policy under which the average holding cost is finite). The ability to stabilize the system is another obvious benefit of using LEWC instead of strict priority policies, such as $c\mu$, which do not belong to the class of stabilizing policies (i.e., policies that always result in a finite cost if the underlying system is stabilizable).

Theorem 5 (Stability under LEWC) *If the condition of Theorem 2 or Corollary 1 is satisfied (and hence, the system is stabilizable), then implementing the LEWC policy stabilizes the "W" system. That is, if* $Z^* = \inf_{\pi \in \Pi} Z^\pi < \infty$ *and* Γ *denotes the LEWC policy, then* $Z^\Gamma < \infty$.

The following theorem presents the same property for the policy of serving the Longest Queue (LQ) as well as the policy of implementing the generalized $c\mu$ ($Gc\mu$) rule with quadratic holding cost.

Theorem 6 (Stability under LQ and $Gc\mu$**)** *Suppose the condition of Theorem 2 or Corollary 1 is satisfied (and hence, the system is stabilizable). Then implementing either the LQ policy or the* $Gc\mu$ *rule with quadratic holding costs stabilizes the "W" system. That is, if* $Z^* = \inf_{\pi \in \Pi} Z^\pi < \infty$, *and* ν *denotes either of these policies policy, then* $Z^\nu < \infty$.

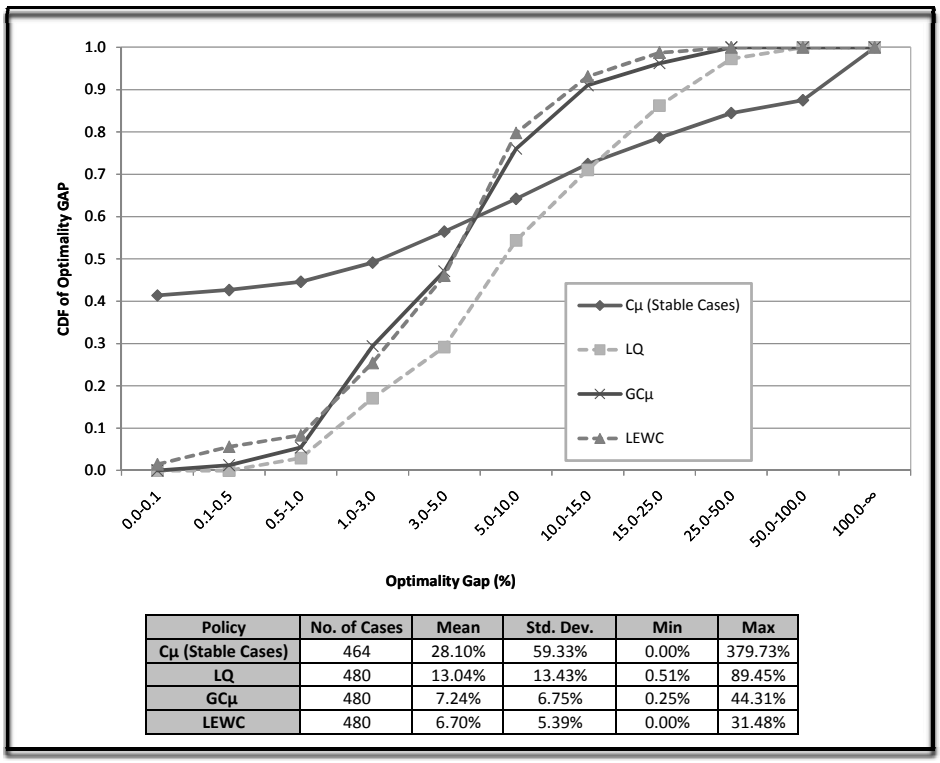
4.4.4 Computational Results

This section compares the performance of our proposed heuristic with (1) the optimal policy, (2) the widely used policy of serving the Longest Queue (LQ), (3)

the well-known $c\mu$ rule, and (4) the generalized $c\mu$ ($Gc\mu$) rule for quadratic holding costs. Under LQ, each server prioritizes serving the queue (among its skill set) with the highest queue length. The $c\mu$ rule, as mentioned before, prescribes server j to serve the queue $k = \text{argmax}_i\, c_i\, \mu_{ji}$, where c_i is the holding cost of a customer in class i. Under the $Gc\mu$, the class to be served by server j is $k = \text{argmax}_i\, \mu_{ji}\, C_i'(x_i(t))$, where $x_i(t)$ is the queue length of class i at time t and $C_i'(\cdot)$ is the derivative of the holding cost function with respect to x_i. As is prevalent in the literature, we use this policy for the case of quadratic holding cost ($C_i(x_i) = c_i\, x_i^2$). Thus, the implemented version of the $Gc\mu$ (also referred as Max-Weight) prescribes server j to serve class $k = \text{argmax}_i\, c_i\, \mu_{ji}\, x_i$. When there is only one job in the shared queue and none in other queues, under all policies we assume the server that is (among available servers) faster in serving the shared task serves the only job in the system.

To perform the comparisons, we developed an extensive test suite of problem instances that covers various combinations of holding costs, disruption rates, service rates, arrival rates, workload distribution among the queues, and system congestions around 70% and 90% (which are common in small service centers and make-to-order manufacturing systems). Part II of the Online Appendix C presents this test suite and the methods used to cover wide ranges of parameter combinations. This test suite generates 480 problem instances for the "W" network and builds a fairly large test suite (given the computational effort for these models).

To benchmark the "W," we employed the Markov Decision Process (MDP) of Section 4.3.2 to compute the optimal cost for each of our problem instances. A similar computational framework is used for *policy evaluation* to benchmark the performance of the LEWC, LQ, $c\mu$, and $Gc\mu$ policies. We used the value-iteration algorithm to solve MDPs numerically and we truncated the state space so that even for the cases with high utilization, probability of reaching the truncation limit was insignificant. Fig. 4.5 summarizes our computational results over the test suite by depicting the empirical Cumulative Distribution Function (CDF) for the percentage optimality gap



Policy	No. of Cases	Mean	Std. Dev.	Min	Max
Cμ (Stable Cases)	464	28.10%	59.33%	0.00%	379.73%
LQ	480	13.04%	13.43%	0.51%	89.45%
GCμ	480	7.24%	6.75%	0.25%	44.31%
LEWC	480	6.70%	5.39%	0.00%	31.48%

Figure 4.5: Performance of $c\mu$, LQ, $Gc\mu$, and LEWC relative to the optimal policy.

(i.e., the CDF of percentage increase over the optimal cost) of each heuristic policy (i.e., LQ, $c\mu$, $Gc\mu$, and LEWC). Specifically, this figure summarizes the result of our $480 \times 5 = 2400$ MDP-based runs. Fig. 4.5 also presents key statistics of the obtained optimality gaps: mean, standard deviation, minimum, and maximum.

Even though the system can be stabilized (by Corollary 1) for each problem instance, we observed that the greedy $c\mu$ policy is unstable in 16 out of 480 problem instances (i.e., 3.33% of cases) within the test suite. Hence, we considered the remaining 464 cases as the basis for computing the statistics on the $c\mu$ rule. However, as Theorems 5 and 6 indicate, LEWC, LQ, and $Gc\mu$ always stabilize the "W." Fig. 4.5 illustrates that the proposed heuristic, LEWC, outperforms the other policies. The mean optimality gap for LEWC is 6.70% in contrast to 13.04% for LQ, 28.10% for $c\mu$ (among stable cases), and 7.24% for $Gc\mu$. That is, the mean optimality gap of LQ, $c\mu$, and $Gc\mu$ are 195%, 420%, and 108% of that of LEWC, respectively. These

results suggest that LEWC (as the first best) and $Gc\mu$ (as the second best) are *nearly optimal* policies considering that the problem instances include *wide variations* on disruption rates, repair rates, arrival rates, costs, traffics, etc. This observation is especially important in light of the following:

- The optimal policy is too complex for practical application in many settings.
- Even for small systems with few servers and task types, obtaining the optimal policy becomes quickly intractable, especially when disruptions are allowed. Therefore, when the size of the systems increases, the optimality gap of a heuristic quickly becomes intractable, so comparisons to the performance of other available heuristics are appropriate.

The standard deviation column in Fig. 4.5 shows that LEWC is considerably more *robust* than other policies in the sense that it is more predictably effective over a wide range of model parameters. For any test case, the heuristics employ the true parameters. Thus, robustness for us is not associated with model uncertainty; rather the range of parameters over which a policy is effective. Indeed, as the figure shows, the standard deviations of LQ, $c\mu$, and $Gc\mu$ are 249%, 1100%, and 125% of that of LEWC, respectively. From Fig 4.5 we also observe that the CDF of the optimality gap of LEWC is closer to that of $Gc\mu$ compared to LQ and $c\mu$. However, LEWC outperforms all of the policies including $Gc\mu$ in all four metrics (mean, standard deviation, min, and max). Moreover, the obtained CDF for the optimality gap of LEWC is *always* above that of LQ, illuminating the clear advantage of using LEWC over LQ. However, the CDF for the optimality gap of $c\mu$ is *initially* above LEWC, because for about 40% of the test problems within our test suite the $c\mu$ rule obtains the optimal cost (since its optimality conditions presented in Theorem 4 are met). Of course, one can revise the LEWC policy so that it implements the $c\mu$ rule when its optimality conditions are met. We did not implement this obvious improvement, because the LEWC policy as stated can be applied in any general network structure

for which the optimality conditions of $c\mu$ may not be known. This way, we gain more confidence that LEWC is suitable for a wide range of applications.

Although the CDF for the optimality gap of $c\mu$ is initially above LEWC, it should be noted that the $c\mu$ rule is a greedy policy and is very risky to implement unless the system's manager can ensure that its optimality conditions are not violated in advance. For instance, even under a heavy-traffic regime, [102] discusses that although the $Gc\mu$ rule is asymptotically optimal when holding costs are convex, its special case, $c\mu$, may not be optimal when the holding costs are linear. Our results for systems with moderate traffic and linear holding costs shows that the $c\mu$ rule performs poorly on average. Moreover, as Fig. 4.5 shows, $c\mu$ is unstable in 3.33% of cases, and among the stable cases, $c\mu$ shows a large standard deviation of 59.33% (and a maximum of 379.73%) in its optimality gap. Our proposed algorithm, similar to $Gc\mu$, combines the cost minimization intuition behind the $c\mu$ and the load balancing idea of LQ. However, unlike $Gc\mu$, LEWC uses an LP (LP 2) to approximate the effort levels (y_{ji}). This use of an LP permits a more accurate estimatation of the workload and allows LEWC to dynamically balance the *workload costs*. This fact makes LEWC not only a more effective policy in terms of the mean optimality gap, but also a considerably more *robust* policy with a relatively small standard deviation of 5.39%. This small standard deviation suggests another advantage of LEWC; using LEWC for comparing various queueing designs (where the optimal policy is computationally intractable) can be more reliable than implementing other policies (see, for instance, [81] and [80] where LQ is used for strategic design comparisons.)

Another observation from Fig. 4.5 is that the widely used LQ and $Gc\mu$ policies are never optimal within our test suite, showing a minimum optimality gap of 0.51% and 0.25%, respectively. However, our proposed LEWC algorithm achieves the optimal cost in a few cases and, like $c\mu$, has a minimum gap of 0%. Moreover, LEWC, unlike $c\mu$ and LQ, rarely results in an optimality gap of above 15%. Indeed, under LEWC the chance of obtaining a performance which is 15% worse than optimal is only 6.9%

(within our test suite), but under LQ and $c\mu$ the chances are 29.0% and 27.6%, respectively.

Another point of interest is to look at the performances of the policies in detail from the perspectives of disruption, congestion, and cost. Table 4.1 presents the detailed comparisons based on Settings (I)-(IV) (see Table 4.6 in Online Appendix C). These four settings represent various combinations of disruption and system congestion. Setting (I) represents a system with no disruption and relatively high traffic. The scope of this research is not the heavy-traffic regime; therefore, here high traffic means relatively high congestion of around 90% and low represents 70%. In Setting (II) the servers are reliable, but the system congestion is relatively low. Settings (III) and (IV) represent scenarios where the system is under relatively lower traffic; in Setting (III), servers are completely reliable, but they are under stochastic disruptions in Setting (IV). The results in Table 4.1 suggest the following observations.

- Interestingly, $c\mu$ outperforms LQ on average under relatively low traffic (see the mean optimality gaps under Settings (II) and (IV)). Under relatively higher traffic, however, LQ is better than $c\mu$. This observation may suggest that the load balancing of LQ becomes more important than the greedy cost minimization of $c\mu$ when traffic is moderate to relatively high.

- LQ is always worse than $c\mu$ with respect to the minimum optimality gap criterion and always better with respect to the maximum optimality gap. This result is intuitive since $c\mu$, unlike LQ, is an extreme (and a greedy) policy. Additionally, LEWC is almost as good as $c\mu$ (which itself outperforms LQ) under the minimum optimality gap criterion and always better than LQ (which itself outperforms $c\mu$) under the maximum optimality gap criterion. Moreover, in all of these four settings, LEWC outperforms LQ, $c\mu$, and $Gc\mu$ with respect to the mean optimality gap and, therefore, presents the best policy under various settings. This strength of LEWC derives from the way it accounts for different parameters of the system through the proposed LP 2 incorporated in the LEWC

Table 4.1: Comparison of policies based on the combinations of disruption and the system congestion using the percentage optimality gaps.

Setting	**Disruption**	**Traffic**	**Policy**	**No. of Cases**	**Mean**	**Min**	**Max**
(I)	No	High	$c\mu$ (Stable Cases)	112	69.52%	0.00%	379.73%
			LQ	120	22.12%	1.25%	89.45%
			$Gc\mu$	120	12.47%	0.87%	44.31%
			LEWC	120	11.92%	0.31%	31.48%
(II)	No	Low	$c\mu$(Stable Cases)	120	7.01%	0.00%	36.75%
			LQ	120	10.45%	0.56%	43.56%
			$Gc\mu$	120	4.86%	0.25%	16.51%
			LEWC	120	4.57%	0.00%	12.06%
(III)	Yes	High	$c\mu$ (Stable Cases)	112	32.75%	0.00%	352.13%
			LQ	120	12.47%	1.23%	62.44%
			$Gc\mu$	120	7.77%	0.71%	29.12%
			LEWC	120	7.06%	0.01%	16.19%
(IV)	Yes	Low	$c\mu$ (Stable Cases)	120	6.36%	0.00%	36.95%
			LQ	120	7.11%	0.51%	34.53%
			$Gc\mu$	120	3.84%	0.33%	12.67%
			LEWC	120	3.25%	0.00%	8.03%

Table 4.2: Comparison of policies based on the holding cost settings (Level of cost asymmetry among different classes: (A) Zero, (B) Low, (C) Moderate, (D) High).

Setting	**Cmu (Stable Cases)** Optimality Gap (%)			**LQ** Optimality Gap (%)			**GCmu** Optimality Gap(%)			**LEWC** Optimality Gap(%)		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
(A)	0.00	0.00	0.00	5.03	1.08	11.39	4.28	0.64	10.99	4.41	0.00	14.52
(B)	29.51	0.00	207.18	8.34	0.56	31.30	7.63	0.44	33.19	5.67	0.14	23.81
(C)	40.75	0.00	379.73	12.77	0.51	51.10	8.69	0.61	44.31	7.17	0.11	29.10
(D)	23.33	0.00	182.31	20.67	1.24	89.45	6.37	0.25	23.39	8.03	0.04	31.48
Total	28.10	0.00	379.73	13.04	0.51	89.45	7.24	0.25	44.31	6.70	0.00	31.48

index.

- All of the policies show a smaller average optimality gap under lower congestion (compare Setting (I) with (II), and Setting (III) with (IV)). This observation may suggest that it is better to implement these policies for systems with low to moderate congestion rather than systems with relatively high traffic.

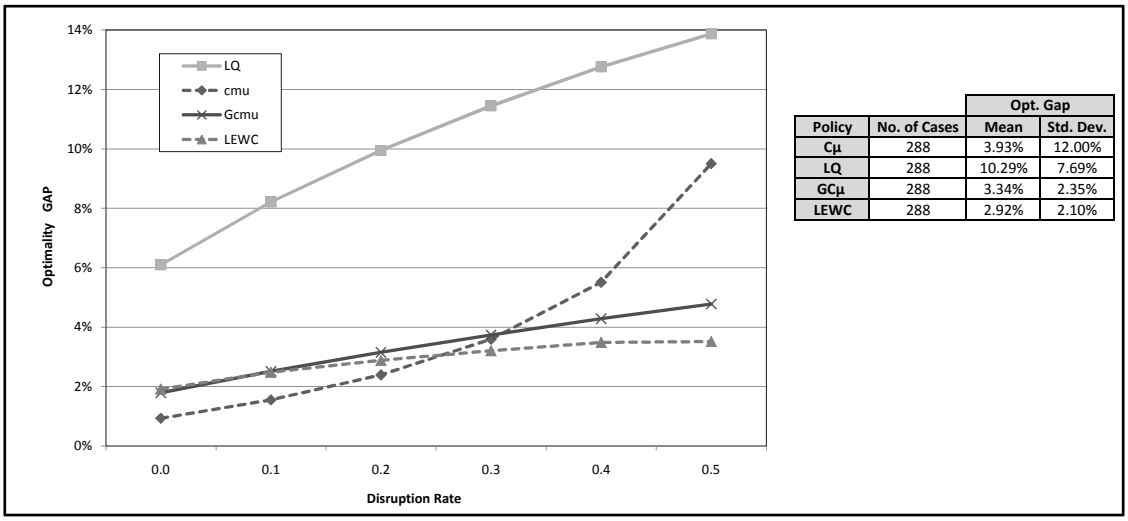
Table 4.2 compares the policies based on the various holding cost settings defined in Table 4.7 in Online Appendix C. In Setting (A), all holding costs equal one, representing a symmetric situation. Settings (B)-(D) represent situations with asymmetric holding costs among customer classes where the degree of asymmetry develops from a low degree in (B) to a high degree in (D). A closer look at Table 4.2 provides the following observations.

- All the policies perform their best (based on the mean criterion) when there is no cost asymmetry. Moreover, the performance of both LQ and LEWC deteriorates as the level of asymmetry increases. However, this deterioration does not occur for $c\mu$ or $Gc\mu$. In fact, both $c\mu$ and $Gc\mu$ perform their worst when the level of asymmetry in holding costs is moderate (Setting (C)). Although the performance of LEWC, unlike $c\mu$ and $Gc\mu$, deteriorates as the level of asymmetry increases, as Table 4.1 showed, LEWC still outperforms both $c\mu$ and $Gc\mu$.

- The proposed heuristic (LEWC) is much more robust to changes in holding costs than other policies. For instance, the mean optimality gap of LQ changes from 5.03% to 20.67% (more than 410% change) by moving from no asymmetry in costs to high asymmetry in costs whereas the mean optimality gap of LEWC only changes from 4.41% to 8.43% (less than 183% change). These results show that the performance of the widely used $c\mu$ and LQ policies, unlike LEWC and $Gc\mu$, is very sensitive to the holding costs. This observation is intuitive, because LQ does not consider holding costs and $c\mu$ depends on them in a relatively extreme way.

- To complete the previous observation, we should note that LEWC, similar to $c\mu$ and $Gc\mu$, uses holding costs to determine the switching curves, but LEWC, unlike $c\mu$ and $Gc\mu$, uses holding cost together with other system parameters. This fact makes LEWC less sensitive to changes in the system parameters (including holding costs and service rates). Therefore, our results recommend the use of LEWC rather than $c\mu$ and $Gc\mu$ when the system parameters vary over time. A similar comparison indicates that LEWC is also preferable to LQ. However, it should be noted that LQ is the rational choice in the absence of any information on the system parameters (which is perhaps its best feature).

Finally, to explore the effect of server disruptions on the performance of LEWC,



		Opt. Gap	
Policy	No. of Cases	Mean	Std. Dev.
Cμ	288	3.93%	12.00%
LQ	288	10.29%	7.69%
GCμ	288	3.34%	2.35%
LEWC	288	2.92%	2.10%

Figure 4.6: Sensitivity of $c\mu$, LQ, $Gc\mu$, and LEWC to variations in disruption risks.

LQ, $c\mu$, and $Gc\mu$, we consider the test suite presented in Tables 4.8, 4.9, and 4.10 (Online Appendix C), and depict in Fig. 4.6 the average optimality gap (over all problem instances) of each of these policies for different levels of disruptions. As the results confirm, LEWC is the least sensitive policy to variations in the disruption risks as it explicitly incorporates disruption rates. This robustness to disruptions, provides another benefit of the proposed policy, LEWC. Among other policies, $Gc\mu$ and LQ are more robust to disruptions compared to $c\mu$, since they implicitly incorporate the effect of disruptions through a consideration of queue lengths.

4.5 Conclusion

This chapter considered the problem of assigning servers to various jobs in real-time to obtain good performance and thereby extract the most benefit from the flexibility of the servers. We first developed a Markov Decision Process (MDP) modeling framework for parallel queueing systems with arbitrary number of job types, arbitrary number of servers, arbitrary flexibility structures, and heterogeneous servers subject to stochastic disruptions. We implemented a Linear Program (LP) to investigate the

stability of such a queueing system, provided some convergence and monotonicity results, and showed that it is sufficient to restrict attention to the class of non-idling stationary policies.

To gain more insights into the characteristics of effective real-time server assignment mechanisms, we then considered the "W" design in which servers are trained to work on a shared task in addition to their fixed task. As a three-class network with two partially flexible servers, the "W" generalizes the "N" structure which has received considerable attention (mainly due to the intrinsic difficulties of the underlying control problem). Next, comparing the "W" design with other possible structures, we showed that the "W" queueing network is an efficient paradigm; with a little investment in flexibility, it provides performance almost as good as a fully flexible network. This chapter provided specific observations into how the "W" design is the preferred structure for many systems with three demand types and two servers. Given the obstacles of fully cross-training servers in practice, our models, analyses, and numerical studies provide designers with a better understanding of *how to effectively introduce limited flexibility to a system.*

We next provided insights for a system manager on *how to benefit from the limited flexibility of servers in real-time.* We showed that, for the "W," a version of the greedy $c\mu$ policy that prescribes every server (whenever not disrupted) to work on *the fixed task before the shared task* is optimal under some conditions. In general, we observed that the optimal policy is of a state-dependent threshold type and can be characterized by four threshold surfaces. However, our findings confirmed that the optimal policy is complex in general, which makes it less attractive for use in practice.

Therefore, we introduced a new, powerful, and implementable policy to control the servers. [137] states that *"A fundamental understanding of the scheduling tradeoff [between achieving server load balancing and scheduling jobs where they are processed most efficiently] is of great theoretical interest."* Our heuristic policy, LEWC, considers this trade-off and balances the *workload cost* of queues (using the traditional notion

of instantaneous workload). This balance is achieved by combining the *load balancing* intuition that is effectively captured in the *queue length dependence* of the LQ policy and the greedy *cost rate minimization* concept embodied in the $c\mu$ rule. LEWC dynamically measures the expected workload costs of the queues, and then assigns each server to the queue with Largest Expected Workload Cost (LEWC) among its skill set. We first established the stability of LEWC (as well as LQ and $Gc\mu$) and then performed an extensive MDP-based numerical test to gain insights into the performance of LEWC as well as three widely used policies: the LQ policy, the $c\mu$, and the $Gc\mu$ rules. Our results particularly suggested the following two conclusions: (1) LEWC is a *near optimal* policy outperforming the other policies, (2) LEWC is more *robust* than LQ, $c\mu$, and $Gc\mu$ over a wide range of operating environments (holding costs, service rates, disruptions, etc.). This latter observation is an important property in practice, since system parameters often vary over time.

This robustness may also suggest that the proposed LEWC heuristic is a more reliable mechanism than the other policies for applications to strategic design, where a designer needs a fair control policy to compare the performance of alternate designs. However, future work could explore if LEWC is also robust across different cross-training designs. If so, it would also be a useful policy for the strategic design of general flexible/queueing systems. This can greatly benefit research targeting strategic design of queueing systems in the vein of [81] and [80].

4.6 Appendix A: Proofs

Proof of Theorem 2 (i) We use the timed round-robin policy of §5.1 of [6] and show that if $\tau^* > 0$, this policy stabilizes the system (although it might not have a good performance). In other words, we show that, if $\tau^* > 0$, this policy results in a *finite* steady-state distribution or equivalently a *finite* performance cost. To show that, consider a special case of the queueing network in [6] with routing matrix $P = 0$

and setups $s_j^{\pi} = 0$, where the long-run availability of server j is given by $\bar{a}_j = \frac{r_j}{\theta_j + r_j}$. Let the optimal decision variables obtained from LP 1 be y_{ji}^*. Also, let λ^* be the optimal objective function of the following LP (used in §5.1 of [6]):

$$\text{Max} \quad \lambda \tag{4.16}$$

subject to:

$$\sum_{j \in \mathcal{N}_s} \delta_{ji}\, \mu_{ji} \geq \lambda \alpha_i \qquad \forall i \in \mathcal{N}_c, \tag{4.17}$$

$$\sum_{i \in \mathcal{N}_c} \delta_{ji} \leq \bar{a}_j \qquad \forall_j \in \mathcal{N}_s, \tag{4.18}$$

$$\delta_{ji} \geq 0 \qquad \forall j \in \mathcal{N}_s, \forall i \in \mathcal{N}_c, \tag{4.19}$$

where $\alpha_i = \frac{\lambda_i}{\sum_{i \in \mathcal{N}_c} \lambda_i}$ is the splitting probability from the current aggregated arrival rate to the system (i.e., $\lambda = \sum_{i \in \mathcal{N}_c} \lambda_i$) to class i. Notice that because y_{ji}^* and $\tau^* > 0$ are feasible for LP (4.8)-(4.11), decision variables $\delta_{ji} = y_{ji}^* \, (\frac{r_j}{\theta_j + r_j})$ and $\hat{\lambda} = (\sum_{i \in \mathcal{N}_c} \lambda_i) + \frac{\tau *}{\alpha_i}$ yield a feasible solution to LP (4.16)-(4.19). Since λ^* is the optimum solution to LP (4.16)-(4.19), we have $\lambda^* \geq \hat{\lambda}$. Also, since $\tau^* > 0$, we have $\hat{\lambda} > \lambda = \sum_{i \in \mathcal{N}_c} \lambda_i$. Hence, $\lambda^* > \lambda$. Now consider the capacity λ (the current aggregated arrival rate to the system) and use the timed round-robin policy of §5.1 of [6]. Following the same line of proof as part (i) of Theorem 3 of that paper and using Theorem 2.4.9 of Dai (1999), we see that the corresponding fluid model of this policy is stable. Hence, using Theorem 4.2 of Dai (1995) and Theorem 4.1 of Dai and Meyn (1995), the underlying Markov process that represents the dynamics of the queueing system under the above-mentioned round-robin policy is positive Harris recurrent and converges to a steady-state distribution with finite moments. This proves part (i).

To prove part (ii), we show that if $\tau^* < 0$, the current aggregated arrival rate to the system ($\lambda = \sum_{i \in \mathcal{N}_c} \lambda_i$) cannot be achieved through LP (4.16)-(4.19). In other words, we show that if $\tau^* < 0$, then $\lambda > \lambda^*$, i.e., the current arrival to the system

is larger than λ^*. The rest of the proof then follows Theorem 3 (ii) of [6]. To show that $\lambda > \lambda^*$ if $\tau* < 0$, we prove the result by contradiction. Suppose $\lambda \leq \lambda^*$ and let δ^*_{ji} be the optimal decision variables of LP (4.16)-(4.19). Then notice that since the variables δ^*_{ji} are feasible to this LP, the set of variables $y_{ji} = \delta^*_{ji}\,(\frac{\theta_j + r_j}{r_j})$ together with $\tau = 0$ provides a feasible solution to LP (4.8)-(4.11). However, since τ^* is the optimum solution to this LP, we will have $\tau^* \geq 0$ which contradicts the assumption $\tau^* < 0$. $\square$

Proof of Theorem 3. In the proof of Theorem (2) we showed that if $\tau^* > 0$, then there exists a policy under which the stochastic process defining the system is positive (Harris) recurrent. Moreover, notice that for any $c > 0$, the set $\{\tilde{\mathbf{X}} : \mathbf{h}\mathbf{X}^T \leq c\}$ is finite. Hence, the assumptions of Corollary 7.5.10 of [126] hold, and from Theorem 7.5.6.(i) of this reference, the set of assumptions [SEN] hold. Now, first notice that the convergence of finite-horizon discounted problem to the infinite-horizon discounted problem follows from Proposition 4.3.1 of [126]. Then, since the set of assumptions [SEN] hold, parts (i), (ii), and (iii) follow from Theorem 7.2.3 of the same reference, if we also notice that $Z^* = \psi Z^*_U$. Part (iv) also follows from Proposition 4.3.1 of [126] for the infinite-horizon problem and then from Theorem 7.2.3 of the same reference for the average-cost case. $\square$

Lemma 4 (Monotonicity) *For any* $\tilde{\mathbf{X}} \in S$, $n \in \mathbb{Z}^+$, $\beta \in [0,1)$, *and* $k \in \mathcal{N}_c$: $V_{n,\beta}\,(\tilde{\mathbf{X}} + \mathbf{e}_k) \geq V_{n,\beta}\,(\tilde{\mathbf{X}})$.

Proof of Lemma 4. We prove this lemma by induction on n. For $n = 0$, we have $V_{0,\beta}\,(\tilde{\mathbf{X}}) = 0 \quad (\forall \tilde{\mathbf{X}} \in S)$. Hence, $V_{0,\beta}\,(\tilde{\mathbf{X}} + \mathbf{e}_k) = V_{0,\beta}\,(\tilde{\mathbf{X}}) = 0\ (\forall \tilde{\mathbf{X}} \in S\,,\ \forall k \in \mathcal{N}_c)$. Now assume for some $n \in \mathbb{Z}^+$ we have $V_{n,\beta}\,(\tilde{\mathbf{X}} + \mathbf{e}_k) \geq V_{n,\beta}\,(\tilde{\mathbf{X}})$ for any $\tilde{\mathbf{X}} \in S$, $\beta \in [0,1)$, and $k \in \mathcal{N}_c$. We show the same monotonicity result is also true for $n+1$.

From (4.5) we have:

$$
\begin{aligned}
&V_{n+1,\beta}(\tilde{\mathbf{X}}+\mathbf{e_k}) = \\
&\mathbf{h}(\mathbf{X}+\mathbf{e_k})^T + \beta \left[\sum_{i\in\mathcal{N}_c} \lambda_i V_{n,\beta}(A_i\tilde{\mathbf{X}}+\mathbf{e}_k) + \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j\, V_{n,\beta}(B_j\tilde{\mathbf{X}}+\mathbf{e}_k) \right. \\
&\qquad\qquad \left. + r_j\,(1-a_j)\, V_{n,\beta}(R_j\tilde{\mathbf{X}}+\mathbf{e}_k)] \right. \\
&+ \min_{\mathbf{u}\in\mathcal{U}(\tilde{\mathbf{X}}+\mathbf{e}_k)} \left\{ \sum_{i\in\mathcal{N}_c}\sum_{j\in\mathcal{N}_s} \mathbb{1}\{u_j=i\}\,\mu_{ji} V_{n,\beta}(D_i\tilde{\mathbf{X}}+\mathbf{e}_k) \right. \\
&\quad \left. \left. + \left(1-\sum_{i\in\mathcal{N}_c}\lambda_i - \sum_{i\in\mathcal{N}_c}\sum_{j\in\mathcal{N}_s}\mathbb{1}\{u_j=i\}\,\mu_{ji} - \sum_{j\in\mathcal{N}_s}[\theta_j\,a_j + r_j\,(1-a_j)]\right) V_{n,\beta}(\tilde{\mathbf{X}}+\mathbf{e}_k) \right\} \right]
\end{aligned}
$$

Now, if the minimization occurs for some control vector $\mathbf{u}^* \in \mathcal{U}(\tilde{\mathbf{X}})$, then after replacing $\mathcal{U}(\tilde{\mathbf{X}}+\mathbf{e}_k)$ with $\mathcal{U}(\tilde{\mathbf{X}})$ the proof would be straightforward since by induction assumption every term in $V_{n+1,\beta}(\tilde{\mathbf{X}}+e_k)$ is greater than or equal to the same term in $V_{n+1,\beta}(\tilde{\mathbf{X}})$ defined in (4.5). However, since $\mathcal{U}(\tilde{\mathbf{X}}+\mathbf{e}_k)$ is a larger admissible set than $\mathcal{U}(\tilde{\mathbf{X}})$, it is not always the case (i.e., $\mathbf{u}^*$ may not belong to $\mathcal{U}(\tilde{\mathbf{X}})$). If $\mathbf{u}^* \notin \mathcal{U}(\tilde{\mathbf{X}})$, writing terms related to the control of class k separately and using the optimal action $\mathbf{u}^*$, we have:

$$
\begin{aligned}
&V_{n+1,\beta}(\tilde{\mathbf{X}}+\mathbf{e}_k) = \\
&\mathbf{h}(\mathbf{X}+\mathbf{e_k})^T + \beta \left[\sum_{i\in\mathcal{N}_c} \lambda_i V_{n,\beta}(A_i\tilde{\mathbf{X}}+\mathbf{e}_k) + \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j\, V_{n,\beta}(B_j\tilde{\mathbf{X}}+\mathbf{e}_k) \right. \\
&\qquad\qquad \left. + r_j\,(1-a_j)\, V_{n,\beta}(R_j\tilde{\mathbf{X}}+\mathbf{e}_k)] \right. \qquad (4.20) \\
&+ \left\{ \sum_{j\in\mathcal{N}_s} \mathbb{1}\{u_j^*=k\}\,\mu_{jk}\,[V_{n,\beta}(\tilde{\mathbf{X}}) - V_{n,\beta}(\tilde{\mathbf{X}}+\mathbf{e}_k)] \right. \\
&\qquad\qquad + \sum_{i\neq k\in\mathcal{N}_c}\sum_{j\in\mathcal{N}_s} \mathbb{1}\{u_j^*=i\}\,\mu_{ji} V_{n,\beta}(D_i\tilde{\mathbf{X}}+\mathbf{e}_k) \\
&\quad \left. \left. + \left(1-\sum_{i\in\mathcal{N}_c}\lambda_i - \sum_{i\neq k\in\mathcal{N}_c}\sum_{j\in\mathcal{N}_s}\mathbb{1}\{u_j^*=i\}\,\mu_{ji} - \sum_{j\in\mathcal{N}_s}[\theta_j\,a_j + r_j\,(1-a_j)]\right) V_{n,\beta}(\tilde{\mathbf{X}}+\mathbf{e}_k) \right\} \right]
\end{aligned}
$$

Now we show that if one uses the same allocation policy as $\mathbf{u}^*$ for every class $i \neq k$ but idles one of the servers (say server j'), which is allocated to class k in $\mathbf{u}^*$, and instead changes $\tilde{\mathbf{X}} + \mathbf{e_k}$ to $\tilde{\mathbf{X}}$, obtains a lower value than $V_{n+1,\beta}(\tilde{\mathbf{X}} + \mathbf{e}_k)$. We will then show that this new value is an upper bound for $V_{n+1,\beta}(\tilde{\mathbf{X}})$ since the new policy (say $\mathbf{u}'$) is admissible (but not necessarily optimal) at state $\tilde{\mathbf{X}}$. Precisely, because we have uniformized the rates

$$\Big(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{i \neq k \in \mathcal{N}_c} \sum_{j \in \mathcal{N}_s} \mathbb{1}\{u_j^* = i\}\, \mu_{ji} - \sum_{j \in \mathcal{N}_s} [\theta_j\, a_j + r_j\, (1 - a_j)] - \mu_{j'k}\Big) \geq 0\,,$$

and thus by induction assumption we have:

$$\Big(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{i \neq k \in \mathcal{N}_c} \sum_{j \in \mathcal{N}_s} \mathbb{1}\{u_j^* = i\}\, \mu_{ji} - \sum_{j \in \mathcal{N}_s} [\theta_j\, a_j + r_j\, (1 - a_j)] - \mu_{j'k}\Big)$$

$$\times \Big[V_{n,\beta}(\tilde{\mathbf{X}} + \mathbf{e}_k) - V_{n,\beta}(\tilde{\mathbf{X}})\Big] \geq 0.$$

Now subtracting this positive term from (4.20), we get:

$$
\begin{aligned}
&V_{n+1,\beta}(\tilde{\mathbf{X}}+\mathbf{e}_k) \geq \\
&\mathbf{h}(\mathbf{X}+\mathbf{e_k})^T + \beta \left[\sum_{i\in\mathcal{N}_c} \lambda_i V_{n,\beta}(A_i\tilde{\mathbf{X}}+\mathbf{e}_k) + \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j\, V_{n,\beta}(B_j\tilde{\mathbf{X}}+\mathbf{e}_k) \right. \\
&\qquad\qquad \left. + r_j\,(1-a_j)\, V_{n,\beta}(R_j\tilde{\mathbf{X}}+\mathbf{e}_k)] \right. \\
&+ \left\{ \sum_{j\neq j'\in\mathcal{N}_s} \mathbb{1}\{u_j^*=k\}\,\mu_{jk}\,[V_{n,\beta}(\tilde{\mathbf{X}}) - V_{n,\beta}(\tilde{\mathbf{X}}+\mathbf{e}_k)] \;+ \sum_{i\neq k\in\mathcal{N}_c}\sum_{j\in\mathcal{N}_s} \mathbb{1}\{u_j^*=i\}\,\mu_{ji} V_{n,\beta}(D_i\tilde{\mathbf{X}}) \right. \\
&\quad \left. \left. + \left(1 - \sum_{i\in\mathcal{N}_c}\lambda_i - \sum_{i\neq k\in\mathcal{N}_c}\sum_{j\in\mathcal{N}_s} \mathbb{1}\{u_j^*=i\}\,\mu_{ji} - \sum_{j\in\mathcal{N}_s}[\theta_j\,a_j + r_j\,(1-a_j)]\right) V_{n,\beta}(\tilde{\mathbf{X}}) \right\} \right] \\
&\geq \mathbf{h}(\mathbf{X})^T + \beta \left[\sum_{i\in\mathcal{N}_c} \lambda_i V_{n,\beta}(A_i\tilde{\mathbf{X}}) + \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j\, V_{n,\beta}(B_j\tilde{\mathbf{X}}) + r_j\,(1-a_j)\, V_{n,\beta}(R_j\tilde{\mathbf{X}})] \right. \\
&+ \left\{ \sum_{j\neq j'\in\mathcal{N}_s} \mathbb{1}\{u_j^*=k\}\,\mu_{jk}\,[V_{n,\beta}(\tilde{\mathbf{X}}) - V_{n,\beta}(\tilde{\mathbf{X}}+\mathbf{e}_k)] \right. \qquad (4.21) \\
&\qquad + \sum_{i\neq k\in\mathcal{N}_c}\sum_{j\in\mathcal{N}_s} \mathbb{1}\{u_j^*=i\}\,\mu_{ji} V_{n,\beta}(D_i\tilde{\mathbf{X}}) \\
&\quad \left. \left. + \left(1 - \sum_{i\in\mathcal{N}_c}\lambda_i - \sum_{i\neq k\in\mathcal{N}_c}\sum_{j\in\mathcal{N}_s} \mathbb{1}\{u_j^*=i\}\,\mu_{ji} - \sum_{j\in\mathcal{N}_s}[\theta_j\,a_j + r_j\,(1-a_j)]\right) V_{n,\beta}(\tilde{\mathbf{X}}) \right\} \right],
\end{aligned}
$$

where the last inequality is obtained by using the induction assumption. Now we show that (4.21) is an upper bound for $V_{n+1,\beta}(\tilde{\mathbf{X}})$, and hence $V_{n+1,\beta}(\tilde{\mathbf{X}}+\mathbf{e}_k) \geq V_{n+1,\beta}(\tilde{\mathbf{X}})$. To see that (4.21) is an upper bound for $V_{n+1,\beta}(\tilde{\mathbf{X}})$, let $\mathbf{u}'$ be a policy that uses the same allocation as $\mathbf{u}^*$ for every class $i \neq k$ but idles server j' (i.e., $u'_{j'} = 0$ and $u'_j = u^*_j$

for every $j \neq j' \in \mathcal{N}_s$). Using $\mathbf{u}'$ and from (4.21), we have:

$$
\begin{aligned}
V_{n+1,\beta}(\tilde{\mathbf{X}} + \mathbf{e}_k) \geq \mathbf{h}\mathbf{X}^T + \beta \Bigg[& \sum_{i \in \mathcal{N}_c} \lambda_i V_{n,\beta}(A_i \tilde{\mathbf{X}}) + \sum_{j \in \mathcal{N}_s} [\theta_j \, a_j \, V_{n,\beta}(B_j \tilde{\mathbf{X}}) \\
& \qquad + r_j \, (1 - a_j) \, V_{n,\beta}(R_j \tilde{\mathbf{X}})] \\
+ \Bigg\{ & \sum_{i \in \mathcal{N}_c} \sum_{j \in \mathcal{N}_s} \mathbb{1}\{u'_j = i\} \, \mu_{ji} V_{n,\beta}(D_i \tilde{\mathbf{X}}) \\
+ & \left(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{i \in \mathcal{N}_c} \sum_{j \in \mathcal{N}_s} \mathbb{1}\{u'_j = i\} \, \mu_{ji} - \sum_{j \in \mathcal{N}_s} [\theta_j \, a_j + r_j \, (1 - a_j)] \right) V_{n,\beta}(\tilde{\mathbf{X}}) \Bigg\} \Bigg] .
\end{aligned}
\tag{4.22}
$$

Finally, notice that since $\mathbf{u}^*$ is admissible at state $\tilde{\mathbf{X}}$, the policy $\mathbf{u}'$ constructed above is admissible (but not necessarily optimal) at state $\tilde{\mathbf{X}}$ based on (4.6). Hence, this policy provides an upper bound for $V_{n+1,\beta}(\tilde{\mathbf{X}})$ defined in (4.5) and the proof is complete. □

Proposition 7 (Non-idling) *There exists an optimal policy (for the finite, infinite, or the average-cost problems) which does not allow for unforced idling.*

Proof of Proposition 7. The following sample path argument shows the result for the finite horizon case. The result for the other cases then follows from Theorem 3. Consider first a $t + n$ period problem. If every optimal policy does not allow for unforced idling, the proof is complete. Otherwise, assume a policy π^* exists that is optimal and allows for unforced idling of a server j. Let t denote the first decision epoch at which π^* allows for unforced idling, i.e., it idles server j although it can assign a class k job to that server. Construct another policy π' that follows the same allocation as π^* until decision epoch t, but at t it assigns server j to a class k job. Since preemption is allowed, at $t + 1$ preempt every server and follow the optimal policy π^*. Let $\tilde{\mathbf{X}}'$ denote the state of the system under policy π'. Notice that until time t both policies have the same cost. However, from $t + 1$, the cost of the policy π' is $V_{n+1,\beta}(\tilde{\mathbf{X}}')$, and the cost of policy π^* is either $V_{n,\beta}(\tilde{\mathbf{X}}' + \mathbf{e}_k)$ or $V_{n+1,\beta}(\tilde{\mathbf{X}}')$.

Hence, using Lemma 1, total cost of policy π' cannot be worse than that of policy π^*. Therefore, since π^* is optimal, π' would be optimal too. Now for the average cost problem notice that:

$$Z^* = \inf_{\pi \in \Pi} \lim_{\beta \to 1^-} \lim_{n \to \infty} \psi(1-\beta) V_{n,\beta}^{\pi}(\tilde{\mathbf{X}}).$$

Hence, since the above argument holds for every $n \in \mathbb{Z}^+$ and $\beta \in (0,1)$, $Z^{\pi'} \leq Z^{\pi^*}$. Therefore, if the policy π^* is optimal with respect to average cost criterion, then policy π' constructed above would be optimal for the average cost problem. □

Proof of Corollary 1. First, LP 1 for the "W" can be written as:

$$\text{Max} \quad \tau \tag{4.23}$$

Subject to:

$$y_{11}\,\mu_{11}\,(\frac{r_1}{\theta_1 + r_1}) \geq \lambda_1 + \tau \tag{4.24}$$

$$y_{23}\,\mu_{23}\,(\frac{r_2}{\theta_2 + r_2}) \geq \lambda_3 + \tau \tag{4.25}$$

$$y_{12}\,\mu_{12}\,(\frac{r_1}{\theta_1 + r_1}) + y_{22}\,\mu_{22}\,(\frac{r_1}{\theta_1 + r_1}) \geq \lambda_2 + \tau \tag{4.26}$$

$$y_{11} + y_{12} \leq 1 \tag{4.27}$$

$$y_{22} + y_{23} \leq 1 \tag{4.28}$$

$$y_{11}, y_{12}, y_{22}, y_{23} \geq 0. \tag{4.29}$$

Let τ^* denote the optimal value of the objective function of this LP. Notice that if $\max\{\rho_1, \rho_3\} > 1$, then either constraint (4.24) (together with (4.27) and (4.29)), or constraint (4.25) (together with (4.28) and (4.29)), forces $\tau^* < 0$. Therefore, by Theorem 2 part (ii) the system is unstable if $\max\{\rho_1, \rho_3\} > 1$. Now suppose $\max\{\rho_1, \rho_3\} < 1$ and $(1-\rho_1)\,\mu_{12}^{eff} + (1-\rho_3)\,\mu_{22}^{eff} > \lambda_2$. Let $\epsilon = 1 - \max\{\rho_1, \rho_3\}$

and $\delta = (1-\rho_1)\,\mu_{12}^{eff} + (1-\rho_3)\,\mu_{22}^{eff} - \lambda_2$. Choose $n \in \mathbb{N}$ large enough such that $\epsilon/n\,(\mu_{12}^{eff} + \mu_{22}^{eff}) < \delta$. Then set $y_{11} = \rho_1 + \epsilon/n$, $y_{12} = 1 - y_{11} = 1 - \rho_1 - \epsilon/n$, $y_{22} = 1 - \rho_3 - \epsilon/n$, and $y_{23} = 1 - y_{22} = \rho_3 + \epsilon/n$. Then notice that, because $\epsilon > 0$ and $\delta > 0$, the set of variables y_{ij} constructed above yield a feasible (but not necessarily optimal) solution for the LP with a positive objective value τ. Since τ^* is the optimal objective value, $\tau^* \geq \tau > 0$. Hence, the system is stable by Theorem 2 part (i), and the proof is complete. □

Proof of Theorem 4 (Non-Collaborative Case). We first prove the case where collaboration is not allowed. We show the result for the finite horizon discounted cost problem defined in (4.5). For the infinite horizon and the average cost criterion, the result would follow from Theorem 3 part (iv). Let $\mathbf{V}$ be the set of real-valued functions on state space S and define functional operators T_h, $T_{\lambda,a}$, T_u, T: $\mathbf{V} \to \mathbf{V}$ as follows:

$$T_h V(\tilde{\mathbf{X}}) = \mathbf{h}\mathbf{X}^T \tag{4.30}$$

$$T_{\lambda,a} V(\tilde{\mathbf{X}}) = \sum_{i \in \mathcal{N}_c} \lambda_i V(A_i \tilde{\mathbf{X}}) + \sum_{j \in \mathcal{N}_s} [\theta_j\, a_j\, V(B_j \tilde{\mathbf{X}}) + r_j\,(1-a_j)\, V(R_j \tilde{\mathbf{X}})] \tag{4.31}$$

$$\begin{aligned} T_u V(\tilde{\mathbf{X}}) = \min_{\mathbf{u} \in \mathcal{U}(\tilde{\mathbf{X}})} \Big\{ & \sum_{i \in \mathcal{N}_c} \sum_{j \in \mathcal{N}_s} \mathbb{1}\{u_j = i\}\, \mu_{ji} V(D_i \tilde{\mathbf{X}}) \\ & + \big(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{i \in \mathcal{N}_c} \sum_{j \in \mathcal{N}_s} \mathbb{1}\{u_j = i\}\, \mu_{ji} \\ & - \sum_{j \in \mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) V(\tilde{\mathbf{X}}) \Big\} \end{aligned} \tag{4.32}$$

$$\begin{aligned} & = \big(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{j \in \mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) V(\tilde{\mathbf{X}}) \\ & - \max_{\mathbf{u} \in \mathcal{U}(\tilde{\mathbf{X}})} \Big\{ \sum_{i \in \mathcal{N}_c} \sum_{j \in \mathcal{N}_s} \mathbb{1}\{u_j = i\}\, \mu_{ji} \Delta_i V(D_i \tilde{\mathbf{X}}) \Big\} \end{aligned} \tag{4.33}$$

$$TV(\tilde{\mathbf{X}}) = T_h V(\tilde{\mathbf{X}}) + \beta \left[T_{\lambda,a} V(\tilde{\mathbf{X}}) + T_u V(\tilde{\mathbf{X}}) \right]. \tag{4.34}$$

From (4.5) and (4.34), and after fixing discount factor β, the N-Period problem satisfies the optimality equation:

$$V_n(\tilde{\mathbf{X}}) = TV_{n-1}(\tilde{\mathbf{X}}) \qquad\qquad n \in \{1, \cdots, N\}, \tag{4.35}$$

with terminal condition $V_0(\cdot) = 0$.
Define $\mathcal{A}_1, \mathcal{A}_2 \subset S$ as:

$$\mathcal{A}_1 = \{\tilde{\mathbf{X}} \in S : a_1 = 1, x_1, x_2 \geq 1\} \text{ and } \mathcal{A}_2 = \{\tilde{\mathbf{X}} \in S : a_2 = 1, x_2, x_3 \geq 1\}. \tag{4.36}$$

Now suppose the condition of the theorem holds. Hence, after uniformization we have $h_1\mu_{11} \geq h_2\mu_{12}$, $h_3\mu_{23} \geq h_2\mu_{22}$, $\mu_{12} \geq \mu_{11}$, and $\mu_{22} \geq \mu_{23}$. First notice that if the shared queue is empty (i.e., if $x_2 = 0$) the result of the theorem is trivial since unforced idling is suboptimal by Proposition 7. Thus, it remains to show that it is optimal (1) for server 1 to serve its fixed task for every $\tilde{\mathbf{X}} \in \mathcal{A}_1$, and (2) for server 2 to serve its fixed task for every $\tilde{\mathbf{X}} \in \mathcal{A}_2$. From (4.33), to show that the optimal allocation in operator T_u follows these rules, it is sufficient to show that the following properties hold:

$$\begin{aligned} &(i)\ \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) \geq \mu_{12}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \quad \text{for all } \tilde{\mathbf{X}} \in \mathcal{A}_1, \\ &(ii)\ \mu_{23}\Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_3) \geq \mu_{22}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \quad \text{for all } \tilde{\mathbf{X}} \in \mathcal{A}_2. \end{aligned} \tag{4.37}$$

Now, let $\mathcal{V} \subset \mathbf{V}$ be the set of real-valued functions such that if $V \in \mathcal{V}$ then V satisfies properties (i) and (ii) given in (4.37). Notice that, since $V_0(\cdot) = 0$, $V_0 \in \mathcal{V}$. Hence, from Lemma 2 (see below) and (4.35) we have $V_n \in \mathcal{V}$ $(\forall n \in \mathbb{Z}^+)$. Thus, the proof is complete for the finite-horizon problem. The proof for infinite-horizon and average cost scenarios then follow from Theorem 3 part (iv). □

Lemma 2 [Preservation (Non-Collaborative Case)]. Suppose that conditions

$$h_1\mu_{11} \geq h_2\mu_{12},\ h_3\mu_{23} \geq h_2\mu_{22},\ \mu_{12} \geq \mu_{11},$$

and $\mu_{22} \geq \mu_{23}$ hold and $V \in \mathcal{V}$. Then $TV \in \mathcal{V}$. That is, operator T preserves properties (i) and (ii) in (4.37) when $h_1\mu_{11} \geq h_2\mu_{12}$, $h_3\mu_{23} \geq h_2\mu_{22}$, $\mu_{12} \geq \mu_{11}$, and $\mu_{22} \geq \mu_{23}$.

Proof. We first show that, under the conditions of this lemma, TV satisfies property (i) in (4.37). To show that, we divide $\mathcal{A}_1$ to five partitions $\mathcal{A}_1^1, \mathcal{A}_1^2, \cdots, \mathcal{A}_1^5$, where

$$\begin{aligned}
\mathcal{A}_1^1 &= \{\tilde{\mathbf{X}} \in \mathcal{A}_1 : x_1 \geq 2, x_2 \geq 1, x_3 \geq 1\},\\
\mathcal{A}_1^2 &= \{\tilde{\mathbf{X}} \in \mathcal{A}_1 : x_1 \geq 2, x_2 \geq 1, x_3 = 0\},\\
\mathcal{A}_1^3 &= \{\tilde{\mathbf{X}} \in \mathcal{A}_1 : x_1 = 1, x_2 \geq 1, x_3 \geq 1\},\\
\mathcal{A}_1^4 &= \{\tilde{\mathbf{X}} \in \mathcal{A}_1 : x_1 = 1, x_2 \geq 2, x_3 = 0\},\\
\mathcal{A}_1^5 &= \{\tilde{\mathbf{X}} \in \mathcal{A}_1 : x_1 = 1, x_2 = 1, x_3 = 0\}.
\end{aligned}$$

Now, for each partition we show that TV satisfies property (i) in (4.37), i.e., we show that $\mu_{11}\Delta_1 TV(\tilde{\mathbf{X}} - \mathbf{e}_1) \geq \mu_{12}\Delta_2 TV(\tilde{\mathbf{X}} - \mathbf{e}_2)$ holds for each partition.

Case 1 ($\tilde{\mathbf{X}} \in \mathcal{A}_1^1$)

Since $V \in \mathcal{V}$, from (4.33) we have:

$$
\begin{aligned}
T_u V(\tilde{\mathbf{X}}) &= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) V(\tilde{\mathbf{X}}) \\
&\quad - \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - a_2\mu_{23}\Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_3) \\
T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) &= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&\quad - \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - 2\mathbf{e}_1) - a_2\mu_{23}\Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_3) \\
T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) &= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
&\quad - \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) - a_2\mu_{23}\Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_2 - \mathbf{e}_3)
\end{aligned}
$$

From above:$\Delta_1 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) =$

$$
\begin{aligned}
&T_u V(\tilde{\mathbf{X}}) - T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&- \mu_{11}\big(\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}} - 2\mathbf{e}_1)\big) - a_2\mu_{23}\big(\Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_3) - \Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_3)\big) \\
&= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&- \mu_{11}\big(\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}} - 2\mathbf{e}_1)\big) - a_2\mu_{23}\big(\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_3)\big),
\end{aligned}
$$

where the last equality is obtained since $\Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_3) - \Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_3) = \Delta_3\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_3) = \Delta_1\Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_3) = \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_3)$.

Similarly, $\Delta_2 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) =$

$$
\begin{aligned}
&T_u V(\tilde{\mathbf{X}}) - T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) = \\
&\big(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{j \in \mathcal{N}_s} [\theta_j \, a_j + r_j \, (1 - a_j)]\big) \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
&- \mu_{11}\big(\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)\big) \\
&\qquad\qquad - a_2 \mu_{23}\big(\Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_3) - \Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_2 - \mathbf{e}_3)\big) \\
&= \big(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{j \in \mathcal{N}_s} [\theta_j \, a_j + r_j \, (1 - a_j)]\big) \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
&- \mu_{11}\big(\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_2) - \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)\big) \\
&\qquad\qquad - a_2 \mu_{23}\big(\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) - \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2 - \mathbf{e}_3)\big).
\end{aligned}
$$

where the last equality is obtained since $\Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_3) - \Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_2 - \mathbf{e}_3) = \Delta_3 \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2 - \mathbf{e}_3) = \Delta_2 \Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_2 - \mathbf{e}_3) = \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) - \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2 - \mathbf{e}_3)$. Therefore, from the above and after simplification we have:

$$
\begin{aligned}
&\mu_{11} \Delta_1 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \mu_{12} \Delta_2 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) = \\
&\big(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{j \in \mathcal{N}_s} [\theta_j \, a_j + r_j \, (1 - a_j)] - \mu_{11} - a_2 \mu_{23}\big) \\
&\qquad\qquad \times \big[\mu_{11} \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \mu_{12} \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2)\big] \\
&+ \mu_{11}\big[\mu_{11} \Delta_1 V(\tilde{\mathbf{X}} - 2\mathbf{e}_1) - \mu_{12} \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)\big] \\
&+ a_2 \mu_{23}\big[\mu_{11} \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_3) - \mu_{12} \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2 - \mathbf{e}_3)\big] \\
&\geq 0,
\end{aligned}
$$

where the inequality holds, since every term is nonnegative as $V \in \mathcal{V}$ (and hence V satisfies property (i)). Thus, operator T_u preserves property (i) in this case. (It should be clear that the coefficient of the first term is nonnegative due to uniformization). Also, notice that clearly operator $T_{\lambda,a}$ preserves this property as well. Moreover, $\mu_{11} \Delta_1 T_h V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \mu_{12} \Delta_2 T_h V(\tilde{\mathbf{X}} - \mathbf{e}_2) = h_1 \mu_{11} - h_2 \mu_{12} \geq 0$. Hence, all of the

operators T_h, $T_{\lambda,a}$, and T_u preserve property (i). Thus, operator $T = T_h + \beta[T_{\lambda,a} + T_u]$ preserves property (i) in (4.37).

Case 2 ($\tilde{\mathbf{X}} \in \mathcal{A}_1^2$)

Since $V \in \mathcal{V}$, for this case from (4.33) we have:

$$\begin{aligned}
T_u V(\tilde{\mathbf{X}}) &= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) V(\tilde{\mathbf{X}}) \\
&\quad - \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - a_2\mu_{22}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) &= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&\quad - \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - 2\mathbf{e}_1) - a_2\mu_{22}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) \\
T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) &= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
&\quad - \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) - a_2\mu_{22}\mathbb{1}\{x_2 \geq 2\}\Delta_3 V(\tilde{\mathbf{X}} - 2\mathbf{e}_2)
\end{aligned}$$

Therefore, $\Delta_1 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) =$

$$\begin{aligned}
&T_u V(\tilde{\mathbf{X}}) - T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big)\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&- \mu_{11}\big(\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}} - 2\mathbf{e}_1)\big) - a_2\mu_{22}\big(\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) - \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)\big) \\
&= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big)\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&- \mu_{11}\big(\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}} - 2\mathbf{e}_1)\big) - a_2\mu_{22}\big(\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)\big),
\end{aligned}$$

where the last equality is obtained since $\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) - \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) = \Delta_2\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) = \Delta_1\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) = \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)$.

Also,

$$\begin{aligned}
\Delta_2 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) &= T_u V(\tilde{\mathbf{X}}) - T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
&= \big(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{j \in \mathcal{N}_s} [\theta_j \, a_j + r_j \, (1 - a_j)]\big) \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
&\quad - \mu_{11} \big(\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)\big) \\
&\quad - a_2 \mu_{22} \big(\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) - \mathbb{1}\{x_2 \geq 2\} \Delta_2 V(\tilde{\mathbf{X}} - 2\mathbf{e}_2)\big) \\
&= \big(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{j \in \mathcal{N}_s} [\theta_j \, a_j + r_j \, (1 - a_j)]\big) \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
&\quad - \mu_{11} \big(\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) - \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)\big) \\
&\quad - a_2 \mu_{22} \big(\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) - \mathbb{1}\{x_2 \geq 2\} \Delta_2 V(\tilde{\mathbf{X}} - 2\mathbf{e}_2)\big)
\end{aligned}$$

where the last equality is obtained since $\Delta_1 \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) = \Delta_2 \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)$, as introduced above in this case.

Hence, from above and after simplification we have:

$$\begin{aligned}
&\mu_{11} \Delta_1 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \mu_{12} \Delta_2 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) = \\
&\big(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{j \in \mathcal{N}_s} [\theta_j \, a_j + r_j \, (1 - a_j)] - \mu_{11} - a_2 \mu_{22}\big) \\
&\qquad \times \big[\mu_{11} \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \mu_{12} \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2)\big] \\
&+ \mu_{11} \big[\mu_{11} \Delta_1 V(\tilde{\mathbf{X}} - 2\mathbf{e}_1) - \mu_{12} \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)\big] \\
&+ a_2 \mu_{22} \big[\mu_{11} \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) - \mu_{12} \mathbb{1}\{x_2 \geq 2\} \Delta_2 V(\tilde{\mathbf{X}} - 2\mathbf{e}_2)\big] \\
&\geq 0,
\end{aligned}$$

where the inequality holds, since if $\mathbb{1}\{x_2 \geq 2\} = 1$, every term is nonnegative as $V \in \mathcal{V}$ (and hence V satisfies property (i)). Also, if $\mathbb{1}\{x_2 \geq 2\} = 0$, then the first and second term are nonnegative as $V \in \mathcal{V}$, and the last term is nonnegative since $\Delta_1 V(\cdot) \geq 0$ from Lemma 4. Thus, operator T_u preserves property (i) in this case. Also, notice that clearly operator $T_{\lambda,a}$ preserves this property. Moreover,

$\mu_{11}\Delta_1 T_h V(\tilde{\mathbf{X}}-\mathbf{e}_1)-\mu_{12}\Delta_2 T_h V(\tilde{\mathbf{X}}-\mathbf{e}_2)=h_1\mu_{11}-h_2\mu_{12}\geq 0$, and thus, operator T_h preserves property (i). Hence, similar to Case 1, all of the operators T_h, $T_{\lambda,a}$, and T_u preserve property (i). Thus, operator $T=T_h+\beta[T_{\lambda,a}+T_u]$ preserves property (i) in (4.37).

Case 3 ($\tilde{\mathbf{X}}\in\mathcal{A}_1^3$)

Since $V\in\mathcal{V}$, from (4.33) we have:

$$\begin{aligned}
T_uV(\tilde{\mathbf{X}}) &= \big(1-\sum_{i\in\mathcal{N}_c}\lambda_i-\sum_{j\in\mathcal{N}_s}[\theta_j\,a_j+r_j\,(1-a_j)]\big)V(\tilde{\mathbf{X}})\\
&\quad-\mu_{11}\Delta_1V(\tilde{\mathbf{X}}-\mathbf{e}_1)-a_2\mu_{23}\Delta_3V(\tilde{\mathbf{X}}-\mathbf{e}_3)\\
T_uV(\tilde{\mathbf{X}}-\mathbf{e}_1) &= \big(1-\sum_{i\in\mathcal{N}_c}\lambda_i-\sum_{j\in\mathcal{N}_s}[\theta_j\,a_j+r_j\,(1-a_j)]\big)V(\tilde{\mathbf{X}}-\mathbf{e}_1)\\
&\quad-\mu_{12}\Delta_2V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)-a_2\mu_{23}\Delta_3V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_3)\\
T_uV(\tilde{\mathbf{X}}-\mathbf{e}_2) &= \big(1-\sum_{i\in\mathcal{N}_c}\lambda_i-\sum_{j\in\mathcal{N}_s}[\theta_j\,a_j+r_j\,(1-a_j)]\big)V(\tilde{\mathbf{X}}-\mathbf{e}_2)\\
&\quad-\mu_{11}\Delta_1V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)-a_2\mu_{23}\Delta_3V(\tilde{\mathbf{X}}-\mathbf{e}_2-\mathbf{e}_3)
\end{aligned}$$

From above: $\Delta_1T_uV(\tilde{\mathbf{X}}-\mathbf{e}_1)=$

$$\begin{aligned}
&T_uV(\tilde{\mathbf{X}})-T_uV(\tilde{\mathbf{X}}-\mathbf{e}_1)\\
&=\big(1-\sum_{i\in\mathcal{N}_c}\lambda_i-\sum_{j\in\mathcal{N}_s}[\theta_j\,a_j+r_j\,(1-a_j)]\big)\Delta_1V(\tilde{\mathbf{X}}-\mathbf{e}_1)\\
&-\mu_{11}\Delta_1V(\tilde{\mathbf{X}}-\mathbf{e}_1)+\mu_{12}\Delta_2V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)\\
&\qquad-a_2\mu_{23}\big(\Delta_3V(\tilde{\mathbf{X}}-\mathbf{e}_3)-\Delta_3V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_3)\big)\\
&=\big(1-\sum_{i\in\mathcal{N}_c}\lambda_i-\sum_{j\in\mathcal{N}_s}[\theta_j\,a_j+r_j\,(1-a_j)]\big)\Delta_1V(\tilde{\mathbf{X}}-\mathbf{e}_1)\\
&-\mu_{11}\Delta_1V(\tilde{\mathbf{X}}-\mathbf{e}_1)+\mu_{12}\Delta_2V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)\\
&\qquad-a_2\mu_{23}\big(\Delta_1V(\tilde{\mathbf{X}}-\mathbf{e}_1)-\Delta_1V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_3)\big),
\end{aligned}$$

where the last equality is obtained since $\Delta_3\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_3) = \Delta_1\Delta_3 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_3)$, as introduced in Case 1.

Similarly, $\Delta_2 T_u V(\tilde{\mathbf{X}}-\mathbf{e}_2) =$

$$
\begin{aligned}
&T_u V(\tilde{\mathbf{X}}) - T_u V(\tilde{\mathbf{X}}-\mathbf{e}_2)\\
&= \big(1-\sum_{i\in\mathcal{N}_c}\lambda_i - \sum_{j\in\mathcal{N}_s}[\theta_j\, a_j + r_j\,(1-a_j)]\big)\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)\\
&- \mu_{11}\big(\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)\big)\\
&\qquad - a_2\mu_{23}\big(\Delta_3 V(\tilde{\mathbf{X}}-\mathbf{e}_3) - \Delta_3 V(\tilde{\mathbf{X}}-\mathbf{e}_2-\mathbf{e}_3)\big)\\
&= \big(1-\sum_{i\in\mathcal{N}_c}\lambda_i - \sum_{j\in\mathcal{N}_s}[\theta_j\, a_j + r_j\,(1-a_j)]\big)\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)\\
&- \mu_{11}\big(\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2) - \Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)\big)\\
&\qquad - a_2\mu_{23}\big(\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2) - \Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2-\mathbf{e}_3)\big).
\end{aligned}
$$

where the last equality is obtained since $\Delta_3\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2-\mathbf{e}_3) = \Delta_2\Delta_3 V(\tilde{\mathbf{X}}-\mathbf{e}_2-\mathbf{e}_3)$, as introduced in Case 1.

Therefore, from the above and after simplification we have:

$$
\begin{aligned}
&\mu_{11}\Delta_1 T_u V(\tilde{\mathbf{X}}-\mathbf{e}_1) - \mu_{12}\Delta_2 T_u V(\tilde{\mathbf{X}}-\mathbf{e}_2) =\\
&\big(1-\sum_{i\in\mathcal{N}_c}\lambda_i - \sum_{j\in\mathcal{N}_s}[\theta_j\, a_j + r_j\,(1-a_j)] - \mu_{11} - a_2\mu_{23}\big)\\
&\qquad\qquad \times \big[\mu_{11}\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1) - \mu_{12}\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)\big]\\
&+ a_2\mu_{23}\big[\mu_{11}\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_3) - \mu_{12}\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2-\mathbf{e}_3)\big]\\
&\geq 0,
\end{aligned}
$$

where the inequality holds, since every term is nonnegative as $V \in \mathcal{V}$ (and hence V satisfies property (i)). Thus, operator T_u preserves property (i) in this case. Also, it is trivial that operator $T_{\lambda,a}$ preserves this property. Moreover, $\mu_{11}\Delta_1 T_h V(\tilde{\mathbf{X}}-\mathbf{e}_1) - \mu_{12}\Delta_2 T_h V(\tilde{\mathbf{X}}-\mathbf{e}_2) = h_1\mu_{11} - h_2\mu_{12} \geq 0$. Hence, similar to the previous cases, all of the

operators T_h, $T_{\lambda,a}$, and T_u preserve property (i). Thus, operator $T = T_h + \beta[T_{\lambda,a} + T_u]$ preserves property (i) in (4.37).

Case 4 ($\tilde{\mathbf{X}} \in \mathcal{A}_1^4$)

Since $V \in \mathcal{V}$, for this case from (4.33) we have:

$$
\begin{aligned}
T_u V(\tilde{\mathbf{X}}) &= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) V(\tilde{\mathbf{X}}) \\
&\quad - \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - a_2\mu_{22}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) &= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&\quad - \mu_{12}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) - a_2\mu_{22}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) \\
T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) &= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
&\quad - \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) - a_2\mu_{22}\Delta_2 V(\tilde{\mathbf{X}} - 2\mathbf{e}_2)
\end{aligned}
$$

Therefore, $\Delta_1 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) =$

$$
\begin{aligned}
&T_u V(\tilde{\mathbf{X}}) - T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big)\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&- \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) + \mu_{12}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) \\
&\qquad - a_2\mu_{22}\big(\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) - \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)\big) \\
&= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big)\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&- \mu_{11}\big(\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}} - 2\mathbf{e}_1)\big) \\
&\qquad - a_2\mu_{22}\big(\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)\big),
\end{aligned}
$$

where the last equality is obtained since $\Delta_2\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) = \Delta_1\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)$,

as introduced in Case 2.

Similarly,

$$\begin{aligned}
\Delta_2 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) &= T_u V(\tilde{\mathbf{X}}) - T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
&= \big(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{j \in \mathcal{N}_s} [\theta_j \, a_j + r_j \, (1 - a_j)]\big) \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
&\quad - \mu_{11} \big(\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)\big) \\
&\quad - a_2 \mu_{22} \big(\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) - \Delta_2 V(\tilde{\mathbf{X}} - 2\mathbf{e}_2)\big). \\
&= \big(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{j \in \mathcal{N}_s} [\theta_j \, a_j + r_j \, (1 - a_j)]\big) \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
&\quad - \mu_{11} \big(\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) - \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)\big) \\
&\quad - a_2 \mu_{22} \big(\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) - \Delta_2 V(\tilde{\mathbf{X}} - 2\mathbf{e}_2)\big).
\end{aligned}$$

where the last equality is obtained since $\Delta_1 \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) = \Delta_2 \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)$, as introduced in Case 2.

Hence, from above and after simplification we have:

$$\begin{aligned}
&\mu_{11} \Delta_1 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \mu_{12} \Delta_2 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) = \\
&\big(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{j \in \mathcal{N}_s} [\theta_j \, a_j + r_j \, (1 - a_j)] - \mu_{11} - a_2 \mu_{22}\big) \\
&\times \big[\mu_{11} \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \mu_{12} \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2)\big] \\
&+ \mu_{11} \big[\mu_{11} \Delta_1 V(\tilde{\mathbf{X}} - 2\mathbf{e}_1) - \mu_{12} \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)\big] \\
&+ a_2 \mu_{22} \big[\mu_{11} \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) - \mu_{12} \Delta_2 V(\tilde{\mathbf{X}} - 2\mathbf{e}_2)\big] \\
&\geq 0,
\end{aligned}$$

where the inequality holds since every term is nonnegative as $V \in \mathcal{V}$ (and hence V satisfies property (i)). Thus, operator T_u preserves property (i) in this case. Also, notice that clearly operator $T_{\lambda,a}$ preserves this property. Moreover, $\mu_{11} \Delta_1 T_h V(\tilde{\mathbf{X}} - \mathbf{e}_1) - \mu_{12} \Delta_2 T_h V(\tilde{\mathbf{X}} - \mathbf{e}_2) = h_1 \mu_{11} - h_2 \mu_{12} \geq 0$, and thus, operator T_h preserves prop-

erty (i). Hence, similar to the previous cases, all of the operators T_h, $T_{\lambda,a}$, and T_u preserve property (i). Thus, operator $T = T_h + \beta[T_{\lambda,a} + T_u]$ preserves property (i) in (4.37).

Case 5 ($\tilde{\mathbf{X}} \in \mathcal{A}_1^5$)

Since $V \in \mathcal{V}$, from (4.33) we have:

$$\begin{aligned} T_u V(\tilde{\mathbf{X}}) = \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) V(\tilde{\mathbf{X}}) \\ - \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - a_2\mu_{22}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \end{aligned}$$

$$\begin{aligned} T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) = \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\ - \max\{\mu_{12}, a_2\mu_{22}\}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) \\ T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) = \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\ - \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2). \end{aligned}$$

Therefore, $\Delta_1 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) =$

$$\begin{aligned} & T_u V(\tilde{\mathbf{X}}) - T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\ & = \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1-a_j)]\big) \Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\ & - \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) + \max\{\mu_{12}, a_2\mu_{22}\}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) - a_2\mu_{22}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2), \end{aligned}$$

and

$$\begin{aligned}\Delta_2 T_u V(\tilde{\mathbf{X}}-\mathbf{e}_2) &= T_u V(\tilde{\mathbf{X}}) - T_u V(\tilde{\mathbf{X}}-\mathbf{e}_2)\\ &= \big(1-\sum_{i\in\mathcal{N}_c}\lambda_i - \sum_{j\in\mathcal{N}_s}[\theta_j\, a_j + r_j\,(1-a_j)]\big)\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)\\ &\quad - \mu_{11}(\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1) - \Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)) - a_2\mu_{22}\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)\\ &= \big(1-\sum_{i\in\mathcal{N}_c}\lambda_i - \sum_{j\in\mathcal{N}_s}[\theta_j\, a_j + r_j\,(1-a_j)]\big)\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)\\ &\quad - \mu_{11}(\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2) - \Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)) - a_2\mu_{22}\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2).\end{aligned}$$

where the last equality is obtained since $\Delta_1\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2) = \Delta_2\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)$, as introduced in Case 2.

Then, from above and after simplification we obtain:

$$\begin{aligned}&\mu_{11}\Delta_1 T_u V(\tilde{\mathbf{X}}-\mathbf{e}_1) - \mu_{12}\Delta_2 T_u V(\tilde{\mathbf{X}}-\mathbf{e}_2) =\\ &\big(1-\sum_{i\in\mathcal{N}_c}\lambda_i - \sum_{j\in\mathcal{N}_s}[\theta_j\, a_j + r_j\,(1-a_j)] - \mu_{11}\big)\big[\mu_{11}\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1) - \mu_{12}\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)\big]\\ &+ a_2\mu_{22}(\mu_{12}-\mu_{11})\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2) + \mu_{11}\big(\max\{\mu_{12}, a_2\mu_{22}\} - \mu_{12}\big)\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)\\ &\geq 0,\end{aligned}$$

where the inequality holds since all the terms in the last expression is nonnegative. The first term is nonnegative as (1) its coefficient is nonnegative due to uniformization, and (2) $V \in \mathcal{V}$ (and hence V satisfies property (i)). Similarly, the second term is also nonnegative as (1) $\mu_{12} \geq \mu_{11}$ is a condition of the lemma, and (2) $\Delta_2 V(\cdot) \geq 0$ by Lemma 4. Finally, the third term is nonnegative as (1) $\max\{\mu_{12}, a_2\mu_{22}\} - \mu_{12} \geq 0$ always holds (note that this also covers the cases with $a_2 = 0$), and (2) $\Delta_2 V(\cdot) \geq 0$ by Lemma 4.

Therefore, operator T_u preserves property (i). Also, notice that clearly operator $T_{\lambda,a}$ preserves this property. Moreover, $\mu_{11}\Delta_1 T_h V(\tilde{\mathbf{X}}-\mathbf{e}_1) - \mu_{12}\Delta_2 T_h V(\tilde{\mathbf{X}}-\mathbf{e}_2) =$

$h_1\mu_{11} - h_2\mu_{12} \geq 0$, and thus, operator T_h preserves property (i). Hence, similar to the previous cases, all of the operators T_h, $T_{\lambda,a}$, and T_u preserve property (i). Thus, operator $T = T_h + \beta[T_{\lambda,a} + T_u]$ preserves property (i) in (4.37).

From the proofs for Cases 1 to 5 above, operator T preserves property (i) for $\tilde{\mathbf{X}} \in \mathcal{A}_1^j$ $(j = 1, \cdots, 5)$. Hence, T preserves this property for $\tilde{\mathbf{X}} \in \cup_{j=1}^5 \mathcal{A}_1^j = \mathcal{A}_1$. It remains to show that T preserves property (ii) in (4.37) for all $\tilde{\mathbf{X}} \in \mathcal{A}_2$. To prove this, notice that because of the complete symmetry in the "W," the proof would follow exactly the same lines as the proof for preservation of property (i), only after *relabeling* servers and customer classes. (Notice that properties (i) and (ii), and all the conditions of the lemma are symmetric). Therefore, T also preserves property (ii) for all $\tilde{\mathbf{X}} \in \mathcal{A}_2$ and the proof is complete. □

Proof of Theorem 4 (Collaborative Case). When servers can collaborate while serving jobs, majority of the proof of Theorem 4 and Lemma 2 stay the same. The functional operators T_h, $T_{\lambda,a}$, T_u, and T and the optimality equation provided in (4.30) to (4.35) of the Theorem 4 proof stay the same, but in order to allow for collaboration, we modify the set of admissible control actions at $\tilde{\mathbf{X}}$ given in (4.6) as follows:

$$\mathcal{U}(\tilde{\mathbf{X}}) = \Big\{ \mathbf{u} = (u_j \in \mathcal{N}_c \cup \{0\} \text{ s.t. } \forall i \in \mathcal{N}_c : \mathbb{1}\{u_j = i\} \leq a_j \, \mathbb{1}\{i \in \mathcal{S}_j\}, \\ \sum_{j \in \mathcal{N}_s} \mathbb{1}\{u_j = i\} \leq |\mathcal{N}_s| \mathbb{1}\{X_i > 0\}) \Big\}. \tag{4.38}$$

Let $\mathcal{A}_1, \mathcal{A}_2 \subset S$ be as defined in (4.36). Now suppose we have $h_1\mu_{11} \geq h_2\mu_{12}$, $h_3\mu_{23} \geq h_2\mu_{22}$, and server collaboration while serving jobs is allowed. Due to the skill structure of the "W," servers can only collaborate on the shared task jobs. Note that, if the shared queue is empty (i.e., if $x_2 = 0$) the result of the theorem is trivial since unforced idling is suboptimal by Proposition 7. Thus, it remains to show that it is optimal (1) for server 1 to serve its fixed task for every $\tilde{\mathbf{X}} \in \mathcal{A}_1$, and (2) for server 2 to serve

its fixed task for every $\tilde{\mathbf{X}} \in \mathcal{A}_2$. From (4.33), to show that the optimal allocation in operator T_u follows these rules, it is sufficient to show that the following properties hold:

$$\begin{aligned} &(i)\ \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) \geq \mu_{12}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \quad \text{for all } \tilde{\mathbf{X}} \in \mathcal{A}_1, \\ &(ii)\ \mu_{23}\Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_3) \geq \mu_{22}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \quad \text{for all } \tilde{\mathbf{X}} \in \mathcal{A}_2. \end{aligned} \tag{4.39}$$

which are the same properties used in the proofs of Theorem 4 and Lemma 2. Properties (i) and (ii) ensure that the server collaboration is the unique optimal action only when (both servers are available and) $\mathbf{X} = (0, 1, 0)$ (where both servers are assigned to fixed task job) and servers follow the $c\mu$ rule, as stated in the Theorem 4, in the rest of the states. To see this, assume V satisfies properties (i) and (ii). As unforced idling is not optimal, T_u in (4.33) is as follows:

$$\begin{aligned} T_u V(\tilde{\mathbf{X}}) = \big(1 - \sum_{i \in \mathcal{N}_c} \lambda_i - \sum_{j \in \mathcal{N}_s} [\theta_j\, a_j + r_j\,(1 - a_j)]\big) V(\tilde{\mathbf{X}}) \\ - \max_{\mathbf{u} \in \mathcal{U}(\tilde{\mathbf{X}})} \Big\{ \sum_{i \in \mathcal{N}_c} \sum_{j \in \mathcal{N}_s} \mathbb{1}\{u_j = i\}\, \mu_{ji} \Delta_i V(D_i \tilde{\mathbf{X}}) \Big\}, \end{aligned}$$

where (when both servers are available)

$$\max_{\mathbf{u} \in \mathcal{U}(\tilde{\mathbf{X}})} \Big\{ \sum_{i \in \mathcal{N}_c} \sum_{j \in \mathcal{N}_s} \mathbb{1}\{u_j = i\}\, \mu_{ji} \Delta_i V(D_i \tilde{\mathbf{X}}) \Big\} =$$

$$\begin{aligned} \max \Big\{ &\mathbb{1}\{x_2 \geq 1\}\big((\mu_{12} + \mu_{22})\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2)\big), \\ &\mathbb{1}\{x_2 \geq 2\}\big(\mu_{12}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) + \mu_{22}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2)\big), \\ &\mathbb{1}\{x_1, x_3 \geq 1\}\big(\mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) + \mu_{23}\Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_3)\big), \\ &\mathbb{1}\{x_2, x_3 \geq 1\}\big(\mu_{12}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) + \mu_{23}\Delta_3 V(\tilde{\mathbf{X}} - \mathbf{e}_3)\big), \\ &\mathbb{1}\{x_1, x_2 \geq 1\}\big(\mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) + \mu_{22}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2)\big) \Big\} \end{aligned} \tag{4.40}$$

The first and the second options in the maximum function are collaboration and assigning each server to separate shared task jobs (if possible), respectively. Properties (i) and (ii) ensure that the first two actions in the maximum function are never the unique optimal action as long as one of the other server assignments is allowable at that state. Therefore, except for states with $\mathbf{X} = (0, x_2, 0)$ where $x_2 \geq 1$, we can ignore the first two actions and construct an optimal policy. For states $\mathbf{X} = (0, x_2, 0)$ where $x_2 \geq 2$, the first two terms in (4.40) are equal. Therefore, as long as $x_2 \geq 2$, we can construct an optimal policy that ignores the collaboration action. As a result, if V satisfies properties (i) and (ii), then collaboration is only uniquely optimal at $\mathbf{X} = \{0, 1, 0\}$. Now, similar to the proof of the non-collaborative case, let $\mathcal{V} \subset \mathbf{V}$ be the set of real-valued functions such that if $V \in \mathcal{V}$ then V satisfies properties (i) and (ii). Notice that, since $V_0(\cdot) = 0$, $V_0 \in \mathcal{V}$. Hence, from Lemma 3 (see below) and (4.35) we have $V_n \in \mathcal{V}$ $(\forall n \in \mathbb{Z}^+)$. Thus, the proof is complete for the finite-horizon problem. The proof for infinite-horizon and average cost scenarios then follow from Theorem 3 part (iv) (notice that Theorem 3 is proved under a non-collaborations assumption, but is also true for the collaborative case). □

Lemma 3 [Preservation (Collaborative Case)]. Suppose $h_1\mu_{11} \geq h_2\mu_{12}$, $h_3\mu_{23} \geq h_2\mu_{22}$, server collaboration is allowed, and $V \in \mathcal{V}$. Then $TV \in \mathcal{V}$. That is, under these conditions, operator T preserves properties (i) and (ii) in (4.39).

Proof. Note that we can use the same partitioning used in the proof of Lemma 2, and in addition, since $V \in \mathcal{V}$, collaboration is the unique optimal action only when (both servers are available and) $\mathbf{X} = \{0, 1, 0\}$. Hence, the proofs for the first four partitions $\mathcal{A}_1^1$ to $\mathcal{A}_1^4$ will not change. In partition $\mathcal{A}_1^5$ however, $T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) = T_u V(0, 1, 0, a_1, a_2)$ and therefore collaboration is optimal when $a_1 = a_2 = 1$. Below we present the proof of Case 5 $(\tilde{\mathbf{X}} \in \mathcal{A}_1^5)$ for the collaborative case.

Case 5 ($\tilde{\mathbf{X}} \in \mathcal{A}_1^5$)

Under the server collaboration assumption, it is optimal to assign both servers to the fixed task when $\mathbf{X} = (0, 1, 0)$ and $a_1 = a_2 = 1$. Therefore, $T_u V(\tilde{\mathbf{X}})$ and $T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2)$ of Case 5 ($\tilde{\mathbf{X}} \in \mathcal{A}_1^5$) proof for the non-collaborative case stay the same, but (regardless of whether $a_2 = 0$ or $a_2 = 1$) the $\max\{\mu_{12}, a_2\mu_{22}\}$ term of $T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1)$ is replaced with $\mu_{12} + a_2\mu_{22}$. Hence, since $V \in \mathcal{V}$, from (4.33) we have:

$$\begin{aligned}
T_u V(\tilde{\mathbf{X}}) &= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1 - a_j)]\big) V(\tilde{\mathbf{X}}) \\
&\quad - \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) - a_2\mu_{22}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) &= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1 - a_j)]\big) V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&\quad - (\mu_{12} + a_2\mu_{22})\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) \\
T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) &= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1 - a_j)]\big) V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
&\quad - \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2).
\end{aligned}$$

Therefore, $\Delta_1 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) =$

$$\begin{aligned}
&T_u V(\tilde{\mathbf{X}}) - T_u V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1 - a_j)]\big)\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) \\
&- \mu_{11}\Delta_1 V(\tilde{\mathbf{X}} - \mathbf{e}_1) + (\mu_{12} + a_2\mu_{22})\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2) - a_2\mu_{22}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2),
\end{aligned}$$

and

$$\begin{aligned}
\Delta_2 T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) &= T_u V(\tilde{\mathbf{X}}) - T_u V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
&= \big(1 - \sum_{i\in\mathcal{N}_c} \lambda_i - \sum_{j\in\mathcal{N}_s} [\theta_j\, a_j + r_j\,(1 - a_j)]\big)\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) \\
&\quad - \mu_{11}(\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2) - \Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_1 - \mathbf{e}_2)) - a_2\mu_{22}\Delta_2 V(\tilde{\mathbf{X}} - \mathbf{e}_2).
\end{aligned}$$

Then, from above and after simplification we obtain:

$$
\begin{aligned}
&\mu_{11}\Delta_1 T_u V(\tilde{\mathbf{X}}-\mathbf{e}_1)-\mu_{12}\Delta_2 T_u V(\tilde{\mathbf{X}}-\mathbf{e}_2)\\
&=\big(1-\sum_{i\in\mathcal{N}_c}\lambda_i-\sum_{j\in\mathcal{N}_s}[\theta_j\,a_j+r_j\,(1-a_j)]-\mu_{11}\big)\big[\mu_{11}\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1)-\mu_{12}\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)\big]\\
&+a_2\mu_{22}(\mu_{12}-\mu_{11})\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)+a_2\mu_{22}\mu_{11}\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)\\
&=\big(1-\sum_{i\in\mathcal{N}_c}\lambda_i-\sum_{j\in\mathcal{N}_s}[\theta_j\,a_j+r_j\,(1-a_j)]-\mu_{11}-a_2\mu_{22}\big)\\
&\times\big[\mu_{11}\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1)-\mu_{12}\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)\big]\\
&+a_2\mu_{22}\mu_{11}\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1)-a_2\mu_{22}\mu_{12}\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)+a_2\mu_{22}(\mu_{12}-\mu_{11})\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)\\
&+a_2\mu_{22}\mu_{11}\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)\\
&=\big(1-\sum_{i\in\mathcal{N}_c}\lambda_i-\sum_{j\in\mathcal{N}_s}[\theta_j\,a_j+r_j\,(1-a_j)]-\mu_{11}-a_2\mu_{22}\big)\\
&\times\big[\mu_{11}\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1)-\mu_{12}\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)\big]\\
&+a_2\mu_{22}\mu_{11}\big(\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1)-\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)+\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)\big)\\
&=\big(1-\sum_{i\in\mathcal{N}_c}\lambda_i-\sum_{j\in\mathcal{N}_s}[\theta_j\,a_j+r_j\,(1-a_j)]-\mu_{11}-a_2\mu_{22}\big)\\
&\times\big[\mu_{11}\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1)-\mu_{12}\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_2)\big]\\
&+a_2\mu_{22}\mu_{11}\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)\geq 0, \qquad (4.41)
\end{aligned}
$$

where the last term is obtained from $\Delta_1\Delta_2 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)=\Delta_2\Delta_1 V(\tilde{\mathbf{X}}-\mathbf{e}_1-\mathbf{e}_2)$, as introduced in Case 2 in the proof of the non-collaborative case.

Note that the inequality in (4.41) holds since all the terms in the last expression are nonnegative. The first term of (4.41) is nonnegative as (1) its coefficient is nonnegative due to uniformization even when $a_2=1$, and (2) $V\in\mathcal{V}$ (and hence V satisfies property (i)). Similarly, the second term is also nonnegative as $\Delta_1 V(\cdot)\geq 0$ by Lemma 4 (which is true when collaboration is allowed).

Hence, by following the same steps employed at the end of the proof for the non-collaborative case, we can conclude that the operator T preserves properties (i) and

(ii) and the proof is complete. $\square$

Proof of Theorem 5. Let $\tilde{\tau}^*$ and y^*_{ji} denote the optimal objective function and the corresponding solutions of LP 2, respectively. Suppose $\tau^* > 0$ (or equivalently $\tilde{\tau}^* > 0$). Then solutions y^*_{ji} satisfy:

$$\sum_{j \in \mathcal{S}_i^{-1}} y^*_{ji} \, (\frac{r_j}{\theta_j + r_j}) \, \mu_{ji} \geq \lambda_i + \tau^* \qquad \forall i \in \mathcal{N}_c, \tag{4.42}$$

$$\sum_{i \in \mathcal{S}_j} y^*_{ji} \leq 1 \qquad \forall_j \in \mathcal{N}_s, \tag{4.43}$$

$$y_{ji} \geq 0 \qquad \forall j \in \mathcal{N}_s, \forall i \in \mathcal{S}_j. \tag{4.44}$$

Define $\kappa_i = h_i \, [\sum_{j \in \mathcal{S}_i^{-1}} y^*_{ji} \, (\frac{r_j}{\theta_j + r_j}) \, \mu_{ji}]^{-1}$ so that $\mathcal{I}_i(x_i) = \kappa_i \, x_i$, where $\mathcal{I}_i(\cdot)$ is the LEWC index of queue i. Next suppose the system under LEWC, denoted Γ, is unstable (i.e., $Z^\Gamma = \infty$) and, therefore, at least one of the queues does not have finite average queue length. Let $\mathcal{M} \neq \emptyset$ denote the set of such queues. That is, $\mathcal{M} = \{i \in \mathcal{N}_c : L_i^\Gamma = \infty\}$. We prove that LEWC stabilizes the system by showing that $\mathcal{M} = \emptyset$ (which is a trivial contradiction).

To show this, we use a fluid model analysis (see, for instance, [35] and [34]). The goal is to show that the analog fluid model of the system drains in a finite time. To this end, fix the control policy and let $Q_i(t)$ and $T_{j,i}(t)$ denote the length of queue i at time $t > 0$ and the time server j spends actively working on class i up to t, respectively. Next, let $(\bar{Q}_i(t), \bar{T}_{j,i}(t) : i \in \mathcal{N}_c; j \in \mathcal{N}_s)$ denote the *fluid limit* of the system (i.e., limit point of $(q^{-1} Q_i(q\,t), q^{-1} T_{j,i}(q\,t) : i \in \mathcal{N}_c; j \in \mathcal{N}_s)$ when $q \to \infty$). First, assume $\mathcal{M} = \{1\}$, i.e., only queue 1 is unstable. We will show that it results in a contradiction by showing that if $\mathcal{M} = \{1\}$, then (under LEWC) queue 1 will indeed drain in a finite time. An important step is to show that for any time t with $\bar{Q}_1(t) > 0$ the length of queue 1 is decreasing on average (and as a result this queue will be drained in a finite time). To show this rigorously, suppose $\bar{Q}_1(t) > 0$ for some t. From page 20

of [34], every component of the fluid model including $\bar{Q}_1(t)$ is Lipschitz continuous, and thus, (similar to the argument used in the proof of [35] Proposition 4.2 part (vi)) we can choose a subsequence $\{q_n^{-1}Q_1(q_n s), n \geq 1\}$ with $q_n \to \infty$ as $n \to \infty$, such that $q_n^{-1}Q_1(q_n s) \to \bar{Q}_1(s) > 0$ uniformly on a small interval $s \in [t, t+h]$. Therefore, for any $M \in \mathbb{R}^+$ there exists $N \in \mathbb{N}$ such that $Q_1(q_n s) > M$ for all $n \geq N$ and all $s \in [t, t+h]$. Since queue 2 is stable, take $M = \kappa_2/\kappa_1 \sup_{q_n s} Q_2(q_n s)$. This choice of M implies that $\kappa_1 Q_1(q_n s) > \kappa_2 Q_2(q_n s)$ for all $n \geq N$, during interval [t,t+h], server 1 spends all his (available) time on queue 1 under LEWC (because queue 1's LEWC index is larger than that of queue 2). Using the definition of derivative, we have $\frac{d}{du}\bar{T}_{1,1}(u)_{|u=t} = \frac{r_1}{\theta_1+r_1}$. That is in the fluid model the departure rate of queue 1 at time t is $\bar{D}_1(t) = \mu_{11}\frac{d}{du}\bar{T}_{1,1}(u)_{|u=t} = \mu_{11}\frac{r_1}{\theta_1+r_1}$. Moreover, since y_{11}^* satisfies (4.42)-(4.44), $\bar{D}_1(t) = \mu_{11}\frac{r_1}{\theta_1+r_1} \geq y_{11}^*\,\mu_{11}\,(\frac{r_1}{\theta_1+r_1}) \geq \lambda_1 + \tau^*$, where $\tau^* > 0$. In other words, the departure rate is (strictly) greater than the arrival rate (the queue length is decreasing) at any time t with $Q_1(t) > 0$. Hence, based on Theorem 2.4.9 of [34] this queue drains in a finite time. Therefore, from Theorem 4.1 of [35] (which connects the stability of the fluid model with that of the original underlying Markov process), $L_1^\Gamma < \infty$. This is a contradiction with $\mathcal{M} = \{1\}$. Thus, $\mathcal{M} \neq \{1\}$. A similar discussion shows that $\mathcal{M} \neq \{3\}$ (because of the symmetry in the "W", it only requires a change in the notation: consider queue 3 as queue 1, and server 2 as server 1).

Next assume $\mathcal{M} = \{2\}$ i.e., queue 2 (the shared queue) is the only unstable queue. Suppose $\bar{Q}_2(t) > 0$ for some $t > 0$. Using a similar discussion, we can choose a subsequence $\{q_n^{-1}Q_2(q_n s), n \geq 1\}$ with $q_n \to \infty$ as $n \to \infty$, such that $q_n^{-1}Q_2(q_n s) \to \bar{Q}_2(s) > 0$ on a small interval $s \in [t, t+h]$. Therefore, for any $M \in \mathbb{R}^+$ there exists $N \in \mathbb{N}$ such that $Q_2(q_n s) > M$ for all $n \geq N$ and all $s \in [t, t+h]$. Now take $M = \max\{\kappa_1/\kappa_2 \sup_{q_n s} Q_1(q_n s), \kappa_3/\kappa_2 \sup_{q_n s} Q_3(q_n s)\}$ and notice that under LEWC both servers 1 and 2 spend their time on queue 2 during a small interval [t,t+h]. Therefore, we have $\frac{d}{du}[\bar{T}_{1,2}(u) + \bar{T}_{2,2}(u)]_{|u=t} = \sum_{j=1,2}\frac{r_j}{\theta_j+r_j}$. That is, in the fluid

model, the effective departure rate of queue 2 at time t is $\bar{D}_2(t) = \mu_{11}\frac{r_1}{\theta_1+r_1} + \mu_{22}\frac{r_2}{\theta_2+r_2}$. Moreover, since y^*_{12} and y^*_{22} satisfy (4.42)-(4.44), we have $\bar{D}_2(t) \geq y^*_{11}\mu_{11}\frac{r_1}{\theta_1+r_1} + y^*_{22}\mu_{22}\frac{r_2}{\theta_2+r_2} \geq \lambda_2 + \tau^*$, where $\tau^* > 0$. Hence, based on Theorem 2.4.9 of [34], queue 2 drains in a finite time. Therefore, from [35] Theorem 4.1, $L_2^\Gamma < \infty$ which is a contradiction with $\mathcal{M} = \{2\}$.

Next assume $\mathcal{M} = \{1, 2\}$ and suppose $\bar{Q}_2(t) > 0$ for some $t > 0$. Similar to the above, first notice that in the fluid model, server 2 spends all his time on queue 2 during a small interval [t,t+h]. Moreover, we can always choose a sequence $\{q_m, m \geq 1\}$ and $L \in \mathbb{N}$ such that $\kappa_1 Q_1(q_m s) < \kappa_2 Q_2(q_m s)$ for all $m \geq L$ and all s in a small interval $[t, t + \delta]$ (since otherwise after some finite time server 1 would be always serving queue 1 under LEWC and hence queue 1 cannot be unstable). Next, similar to the above and since $Q_2(t)$ is Lipschitz continuous, we can consider a subsequence of this sequence, $\{q_n, n \geq 1\}$, such that $q_n^{-1}Q_2(q_n s) \to \bar{Q}_2(s)$ on a small interval $[t, t+\epsilon]$. Thus, in the fluid model, server 1 is serving queue 2 during an interval $[t, t + \epsilon]$. Therefore, on $[t, t + \min\{h, \epsilon\}]$ both servers are serving queue 2 in the fluid model. Taking limit we have $\frac{d}{du}[\bar{T}_{1,2}(u) + \bar{T}_{2,2}(u)]_{|u=t} = \sum_{j=1,2}\frac{r_j}{\theta_j+r_j}$. Hence, the effective departure rate from queue q at time t is $\bar{D}_2(t) = \mu_{11}\frac{r_1}{\theta_1+r_1} + \mu_{22}\frac{r_2}{\theta_2+r_2}$. Moreover, since y^*_{12} and y^*_{22} satisfy (4.42)-(4.44), we have $\bar{D}_2(t) \geq y^*_{11}\mu_{11}\frac{r_1}{\theta_1+r_1} + y^*_{22}\mu_{22}\frac{r_2}{\theta_2+r_2} \geq \lambda_2 + \tau^*$, where $\tau^* > 0$. Hence, based on Theorem 2.4.9 of [34] queue 2 drains in a finite time. Therefore, from [35] Theorem 4.1, $L_2^\Gamma < \infty$ which is a contradiction. Thus, $M \neq \{1, 2\}$. Moreover, because of the symmetry in "W", a relabeling of the queues shows that $M \neq \{2, 3\}$.

The analysis for the only remaining case, i.e, $M = \{1, 2, 3\}$ also follows a similar line of proof but by choosing appropriate subsequences such that $\kappa_2 Q_2(q_n s) > \max\{\kappa_i Q_i(q_n s), i = 1, 3\}$. This choice will result in the stability of queue 2 which is a contradiction. □

Proof of Theorem 6. For the case of LQ policy, set $\kappa'_i = 1$ and notice that each

server serves the queue with highest $\kappa_i' x_i$ among its skill set. For the $Gc\mu$ rule with quadratic holding cost define $\kappa_{ji}'' = h_i \, \mu_{ji}$ and notice that server j serves the queue with highest $\kappa_{ji}'' x_i$. The proof is then similar to the proof of Theorem 5, after replacing $\kappa_i = h_i \, [\sum_{j \in \mathcal{S}_i^{-1}} \, y_{ji}^* \, (\frac{r_j}{\theta_j + r_j}) \, \mu_{ji}]^{-1}$ of LEWC with κ_i' for the LQ policy and with κ_{ji}'' for the $Gc\mu$ rule. $\square$

4.7 Appendix B: Structure of the Optimal Policy

Our extensive MDP-based numerical computations show that the optimal policy for a "W" structure with server disruptions is complex in general. The following conjecture characterizes the optimal policy as a *state-dependent threshold type policy* which can be defined by four switching surfaces. We state this result as a conjecture and support it through some numerical examples.

Conjecture 1 (Optimality of a State-Dependent Threshold Policy) *Consider a stabilizable "W" with stochastic disruptions (or without them as a degenerate case) and define threshold surfaces from* $\mathbb{N}^{+2}$ *to the extended positive integers as:*

$$g_1^{up}(x_2, x_3) = \inf \left\{ x_1 \in \mathbb{N} : \; \mu_{11} \, \Delta_1 J(x_1 - 1, x_2, x_3, 1, 1) \geq \mu_{12} \, \Delta_2 J(x_1, x_2 - 1, x_3, 1, 1) \right\}$$

$$g_1^{down}(x_2, x_3) = \inf \left\{ x_1 \in \mathbb{N} : \; \mu_{11} \, \Delta_1 J(x_1 - 1, x_2, x_3, 1, 0) \geq \mu_{12} \, \Delta_2 J(x_1, x_2 - 1, x_3, 1, 0) \right\}$$

$$g_2^{up}(x_1, x_2) = \inf \left\{ x_3 \in \mathbb{N} : \; \mu_{23} \, \Delta_3 J(x_1, x_2, x_3 - 1, 1, 1) \geq \mu_{22} \, \Delta_2 J(x_1, x_2 - 1, x_3, 1, 1) \right\}$$

$$g_2^{down}(x_1, x_2) = \inf \left\{ x_3 \in \mathbb{N} : \; \mu_{23} \, \Delta_3 J(x_1, x_2, x_3 - 1, 0, 1) \geq \mu_{22} \, \Delta_2 J(x_1, x_2 - 1, x_3, 0, 1) \right\}$$

with the convention that $\inf \emptyset = \infty$. *Then there exists an optimal state-dependent*

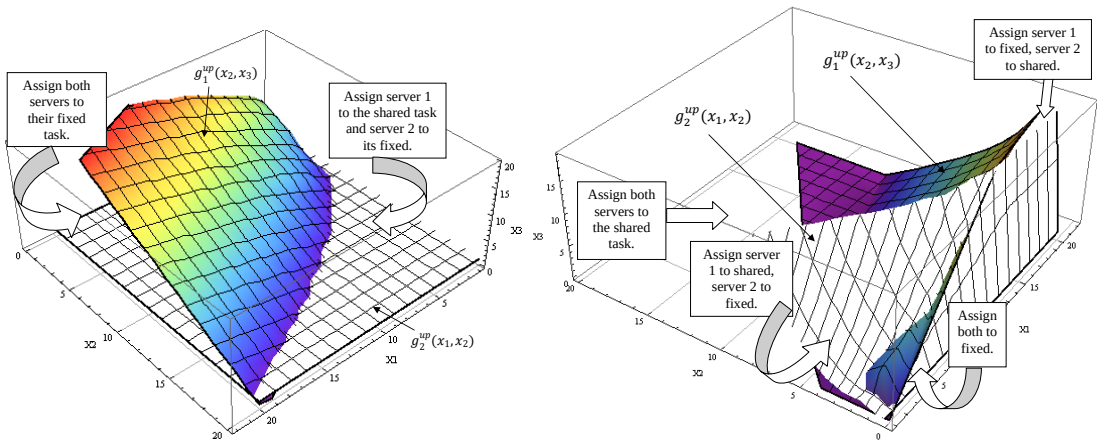
threshold type policy that acts as follows. In each state $\tilde{\mathbf{X}}$*, the optimal policy assigns the working server 1 (server 2) to its fixed task if, and only if,* $x_1 \geq g_1^{up}(x_2, x_3)$ *(*$x_3 \geq g_2^{up}(x_1, x_2)$*) when the other server is up, and* $x_1 \geq g_1^{down}(x_2, x_3)$ *(*$x_3 \geq g_2^{down}(x_1, x_2)$*) when the other server is down.*

The above conjecture states that the four threshold/switching surfaces $g_j^{up}(\cdot)$ and $g_j^{down}(\cdot)$ (for $j = 1, 2$) characterize an optimal policy in general. This conjecture is based on the intuition (observed numerically), that the optimal policy is monotone in the number of jobs in the fixed queues: if it is optimal to assign server 1 (server 2) to its fixed task in state $\tilde{\mathbf{X}}$, then it is also optimal to do so in state $\tilde{\mathbf{X}} + \mathbf{e}_1$ ($\tilde{\mathbf{X}} + \mathbf{e}_3$). Therefore, for every server, it is sufficient to recognize the minimum number of jobs in that server's fixed task such that it is optimal to start serving the fixed task. Such states are defined by the switching surfaces $g_j^{up}(\cdot)$ and $g_j^{down}(\cdot)$ (for $j = 1, 2$).

Remark. Theorem 4 presented a special case of the above conjecture by setting all the surface functions to the value 1. Another special case can be seen in [39] for the so-called "N" structure without disruptions, where the authors show the optimality of a threshold type policy for the case when all servers have an equal service rate (see Theorem 5.4 part 4 of [39]).

The following numerical examples support the above conjecture and clarify the behavior of threshold surfaces by illustrating the optimal policy.

Example 1. Consider a "W" queueing network where servers are reliable (i.e., a system without disruptions) with parameters $\lambda_1' = 0.7$, $\lambda_2' = 0.9$, $\lambda_3' = 0.4$, $\mu_{11}' = \mu_{12}' = 1.0$, $\mu_{22}' = 1.2$, $\mu_{23}' = 1.5$, $h_1' = 1.0$, $h_2' = 1.2$, and $h_3' = 1.0$. Fig. 4.7(a) illustrates the optimal policy. Since there is no disruption, the optimal policy can be described by threshold surfaces $g_1^{up}(x_2, x_3)$ and $g_2^{up}(x_1, x_2)$. As Fig. 4.7(a) shows, server 2 follows the $c\mu$ rule by giving strict priority to its fixed task. Thus, $g_2^{up}(x_1, x_2) = 1$ for every $x_1, x_2 \in \mathbb{Z}^+$. However, for server 1, serving the fixed is optimal if, and only if, $x_1 \geq g_1^{up}(x_2, x_3)$. It should be noted that there are states where servers serve the shared task to avoid idling. However, these states are hidden behind



(a) Example 1. (b) Example 2.

Figure 4.7: Illustration of the optimal policy in examples 1 and 2.

the surface $g_2^{up}(x_1, x_2) = 1$ in Fig. 4.7. Since one of the servers is following a strict priority rule, the optimal policy is still simple and intuitive. However, as the next examples show, the optimal policy can be much more complex.

Example 2. Consider a "W" queueing network where servers are reliable (i.e., a system without disruptions) with parameters $\lambda_1' = 0.7$, $\lambda_2' = 0.9$, $\lambda_3' = 0.4$, $\mu_{11}' = 1.2$, $\mu_{12}' = 1.0$, $\mu_{22}' = 1.0$, $\mu_{23}' = 1.2$, $h_1' = 1.0$, $h_2' = 1.5$, and $h_3' = 1.0$. Fig. 4.7(b) illustrates the optimal policy. As can bee seen in Fig. Fig. 4.7(b), the optimal policy is much more complicated in this case, since both servers use threshold surfaces. These two threshold surfaces divide the state space into four regions. As is shown in Fig. 4.7(b) (using wide arrows), in each region a specific allocation remains optimal. As with all other numerical examples we performed, the observed behavior is consistent with the conjecture made earlier. The next example presents a case where the servers are subject to stochastic disruptions and again adheres to the earlier conjecture.

Example 3. Consider a "W" structure where servers are not reliable and the parameters of the system are: $\lambda_1' = 0.3$, $\lambda_2' = 0.2$, $\lambda_3' = 0.1$, $\mu_{11}' = 1.0$, $\mu_{12}' = 1.2$, $\mu_{22}' = 1.2$, $\mu_{23}' = 1.0$, $h_1' = 1.1$, $h_2' = 1.2$, $h_3' = 1.3$, $\theta_1' = 0.4$, $\theta_2' = 0.45$, $r_1' = 0.6$, and $r_2' = 0.55$. Fig. 4.8 illustrates the optimal policy. In Fig. 4.8, the graph on left depicts $g_1^{up}(x_2, x_3)$ and $g_2^{up}(x_1, x_2)$. The graph on right illustrates $g_1^{down}(x_2, x_3)$

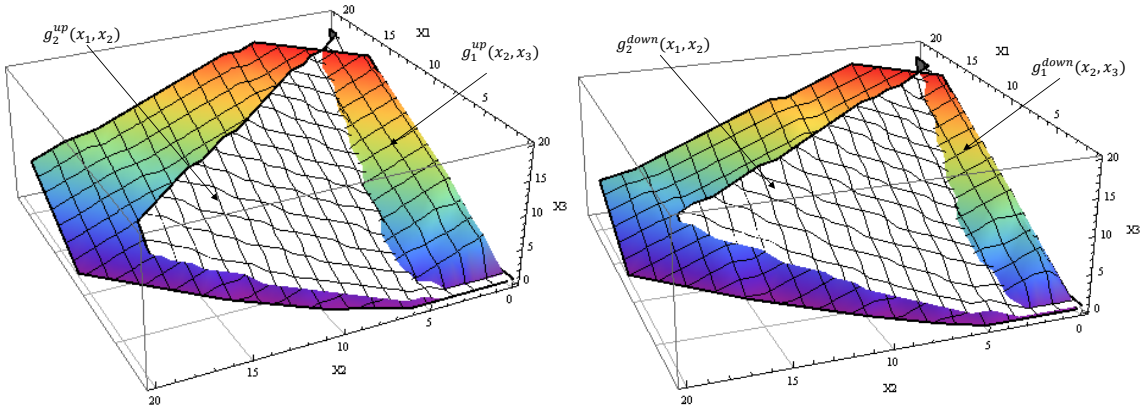


Figure 4.8: Illustration of the optimal policy in example 3.

and $g_2^{down}(x_1, x_2)$. Hence, the left graph shows the optimal policy when both servers are available. The right graph, depicts the optimal behavior of one server when the other server is disrupted. The four threshold surfaces depicted in Fig. 4.8 completely define the optimal policy, and the result is consistent with the earlier conjecture. A first glance at Fig. 4.8 might indicate that the threshold surfaces are the same regardless of the other server's availability (i.e., $g_j^{up}(\cdot) = g_j^{down}(\cdot)$). However, a closer look reveals that the threshold surfaces $g_j^{up}(\cdot)$ and $g_j^{down}(\cdot)$ (for both $j = 1, 2$) are different at many points. For instance, near the origin, $g_1^{down}(\cdot)$ is lower than $g_1^{up}(\cdot)$ (e.g., (2,1,6) is on $g_1^{up}(\cdot)$, but (2,1,1) is on $g_1^{down}(\cdot)$). In this example, to start serving its fixed task, server 1 requires fewer jobs at queue 3 when server 2 is down than when it is is working.

4.8 Appendix C: Test Suites for Numerical Comparisons

This appendix presents the test suites for our numerical studies. Part I describes the test suite for comparing the "W" design with other possible designs and is related to Figures 4.3 and 4.4. Part II presents the test suite for comparing the performance of our proposed Largest Expected Workload Cost (LEWC) heuristic with other well-

known control policies (Section 4.4.4).

4.8.1 PART I: Comparisons of Various Structures (Figures 4.3 and 4.4)

The parameter settings used in Fig. 4.4 are presented in Table 4.3. As is the case throughout the paper, we distinguish between non-uniformized parameters and the uniformized parameters using prime notation. In this table, ρ denotes the congestion factor of the system, which varies from 0.1 to 0.9 as is illustrated in Fig. 4.4. In Table 4.3, as is introduced in the main body, $\mathcal{S}_j$ represents the skill set or capability set of server j. For instance, in Suite 1, server 1 works at a speed of 3 and server 2 works at a speed of 2. To eliminate the homomorphic designs in these test suites, we assume job type two (the shared task) has an arrival rate greater than or equal to the fixed types. Moreover, based on Theorem 2, for large values of ρ in some test suites, some structures with low flexibility can be unstable. This fact is considered in the illustration presented in Fig. 4.4. For instance, as is clear from Fig. 4.4, Structure 1 is not stabilizable for large values of ρ in Suites 1, 2, and 4.

4.8.2 Part II: Comparisons of LEWC with Other Well-Known Policies

Assume that every source (worker, server, machine, agent, etc.) is at least as good as a standard source (which serves in expectation one unit of work content in one unit of time). We consider different skill levels (heterogeneity in service rates) and allow some sources to be 20% faster in working on some job types. Table 4.4 presents various combinations of interest assuming that servers are faster (in a non-strict sense) in serving their specialized (i.e., fixed) task. (Notice that otherwise, there would be another design for which the current shared task is fixed for that rapid server, and the

Table 4.3: Various Suites of Parameter Settings for Comparisons of Alternate Structures.

Parameters	**Suite 1**	**Suite 2**	**Suite 3**	**Suite 4**
λ'_1	$4/3\,\rho$	$0.25\,\rho$	$1\,\rho$	$4/3\,\rho$
λ'_2	$4/3\,\rho$	$3.50\,\rho$	$2\,\rho$	$4/3\,\rho$
λ'_3	$4/3\,\rho$	$0.25\,\rho$	$1\,\rho$	$4/3\,\rho$
$\mu'_{1i}\,(\forall i \in \mathcal{S}_1)$	3.0	2.0	2.0	2.0
$\mu'_{2i}\,(\forall i \in \mathcal{S}_2)$	2.0	2.0	2.0	2.0
θ'_1	0.1	0	0	0
θ'_2	0.1	0	0	0
r'_1	0.9	0	0	0
r'_2	0.9	0	0	0

current fixed task of that rapid server is shared between both servers.) Moreover, for every service rate setting, we consider four workload distributions as shown in Table 4.5 . These workload distributions introduce asymmetry in the amount of the time that each server needs to spend among the queues. Setting (a) in this table represents a case where both servers need to spend their time equally between their fixed task and the shared task. This case represents a symmetric workload distribution among the queues. Settings (b)-(d) consider other extreme workload distributions where a server may need to spend 80% of its time on either the fixed task or the shared task. Notice that (1) because of the inherent symmetry in the "W", these settings cover all key workload distributions with the above assumption, and (2) these workload distributions are approximate and the actual workloads of servers depend on the control policy implemented.

Considering these workload distributions, Table 4.4 also presents the corresponding arrival rates as a function of a congestion factor, ρ. Assuming that both servers are 100% available, a congestion factor of 1 indicates the maximum possible arrival rates to each queue (which corresponds to a system with 100% utilization) based on

the corresponding workload distribution. Hence, by computing the arrival rates in this way: (1) we can ensure that the problem instances are within the feasible/stable region, and (2) by changing ρ, we can control the overall utilization (or traffic) of the system. As shown in Table 4.5 , we choose ρ such that for every setting we can include a system in both 90% (relatively high) and 70% (relatively low) utilization. Notice that the actual system utilization also depends on the control policy implemented. Since there is no analytical method to exactly determine the real utilization of the system, we have reported the approximate system utilization based on the ratio of work content over available capacity as is presented in Table 4.6.

Table 4.6 also presents disruption and repair rates corresponding to combinations of 100% (representing no-disruption scenarios), 90% (representing relatively high availability), and 70% (representing relatively low availability) of servers. Specifically, Settings (I) and (II) in this table illustrate systems with completely reliable servers under relatively high and low traffic, respectively. Settings (III) and (IV) represent systems in which one server is of low reliability and the other server is highly reliable. Since disruption and repair dynamics form a two-state Markov chain, we can precisely compute the steady-state availabilities. In fact, the ratio $r_j/(r_j + \theta_j)$ is the steady-state availability of server j and is reported in Table 4.6 . Moreover, we assume that both servers are repaired through a similar process. Hence, the repair rates are selected to be close. However, by considering different repair rates, we also include cases in which servers have asymmetric reliability. Because of the symmetry in the "W," it is sufficient to consider server 1 as the server that possesses a lower reliability. To capture the effect of disruptions (for scenarios that allow disruptions), repair rates are also selected such that when a server is disrupted, it takes approximately 3 (precisely 1/0.35) standard service durations until the server returns to a working state.

Finally, Table 4.7 presents different holding cost settings. In Setting (A), all job types have a similar holding cost, representing a symmetric case. Settings (B)-(D)

introduce asymmetry in holding costs. More precisely, in Settings (B), (C), and (D) the asymmetry among the costs is obtained by using multipliers of 0.25, 0.5. and 1, respectively. Hence, the degree of asymmetry in holding costs (i.e., differences in costs) increases from Setting (A) to (D). Moreover, costs within each of these settings are chosen such that they cover all the possible $c\mu$ order-based permutations (among the three queues). In fact, since it is sufficient to consider only the highest and the second highest $c\mu$ rankings (as the third one will follow), the last column of Table 4.7 reports the corresponding $c\mu$ ranking permutation among the queues. Because of the symmetry between fixed queues, we can assume the holding cost associated with one of the fixed queues (h_1 or h_3) is always greater than the holding cost associated with the other fixed queue. We assume ($h_3 \geq h_1$) to eliminate the degenerate cases. These cost settings together with service rate settings (Table 4.4) and disruption settings (Table 4.6) generate $10 \times 12 \times 4 = 480$ different problem instances that cover a fair and extensive range of possibilities regarding arrivals, holding costs, disruptions, traffic, service rates, etc.

Table 4.4: Various service rate combinations and arrival rates in the test suites.

Setting	**Service Rates**				**Workload Distr.**	**Arrivals**		
	μ'_{11}	μ'_{12}	μ'_{22}	μ'_{23}		λ'_1	λ'_2	λ'_3
1	1.0	1.0	1.0	1.0	(a)	$0.5\,\rho$	$1\,\rho$	$0.5\,\rho$
2	1.0	1.0	1.0	1.0	(b)	$0.8\,\rho$	$1\,\rho$	$0.2\,\rho$
3	1.0	1.0	1.0	1.0	(c)	$0.8\,\rho$	$0.4\,\rho$	$0.8\,\rho$
4	1.0	1.0	1.0	1.0	(d)	$0.2\,\rho$	$1.6\,\rho$	$0.2\,\rho$
5	1.0	1.0	1.0	1.2	(a)	$0.5\,\rho$	$1\,\rho$	$0.6\,\rho$
6	1.0	1.0	1.0	1.2	(b)	$0.8\,\rho$	$1\,\rho$	$0.24\,\rho$
7	1.0	1.0	1.0	1.2	(c)	$0.8\,\rho$	$0.4\,\rho$	$0.96\,\rho$
8	1.0	1.0	1.0	1.2	(d)	$0.2\,\rho$	$1.6\,\rho$	$0.24\,\rho$
9	1.2	1.0	1.0	1.2	(a)	$0.6\,\rho$	$1\,\rho$	$0.6\,\rho$
10	1.2	1.0	1.0	1.2	(b)	$0.96\,\rho$	$1\,\rho$	$0.24\,\rho$
11	1.2	1.0	1.0	1.2	(c)	$0.96\,\rho$	$0.4\,\rho$	$0.96\,\rho$
12	1.2	1.0	1.0	1.2	(d)	$0.24\,\rho$	$1.6\,\rho$	$0.24\,\rho$

Tables 4.8, 4.9, and 4.10 present the test suite used for investigating the effect of

Table 4.5: Workload distributions on skills used in Table 4.4 (S: Server; C: Class of Customer).

Setting	**Workload Distributions**			
	S1-C1	S1-C2	S2-C2	S2-C3
(a)	50%	50%	50%	50%
(b)	80%	20%	80%	20%
(c)	80%	20%	20%	80%
(d)	20%	80%	80%	20%

Table 4.6: Settings of disruption rates, repair rates, and system congestion factor (ρ) in addition to the corresponding system utilization.

Setting	**Disrupt.&Rep.**				**Availability**		**Conges. (ρ)**	**Util.**
	r_1	r_2	θ_1	θ_2	Server 1	Server 2		
(I)	0	0	0	0	100%	100%	0.9	$\sim 90\%$
(II)	0	0	0	0			0.7	$\sim 70\%$
(III)	0.35	0.36	0.15	0.04	70%	90%	0.63	$< 90\%$
(IV)	0.35	0.36	0.15	0.04			0.49	$< 70\%$

disruptions (Fig. 4.6). Table 4.8 corresponds to Table 4.4 with a congestion factor of $\rho = 0.5$. Table 4.9 presents the various holding cost settings considered, and Table 4.10 presents the settings related to disruptions and repair. Server 1 is considered to be completely reliable. However, disruption rate of Server 2 changes from 0.0 to 0.5, while its repair rate is kept constant to generate different steady-state realisability levels.

Table 4.7: Holding cost settings and the corresponding highest and second highest $c\mu$ rankings among the queues (SF: Highest=Shared, Second Highest=Fixed; FS: Highest=Fixed, Second Highest=Shared, FF: First and Second Highest=Fixed).

Setting	**Cost Asym.**	**Holding Cost**			$c\mu$ **Ranking Permut.**	**Cases**
		h_1	h_2	h_3		
(A)	Zero	1	1	1	Symmetric	48
(B)	Low	1	1.5	1.25	SF	48
		1	1.25	1.5	FS	48
		1.25	1	1.5	FF	48
(C)	Moderate	1	2	1.5	SF	48
		1	1.5	2	FS	48
		1.5	1	2	FF	48
(D)	High	1	3	2	SF	48
		1	2	3	FS	48
		2	1	3	FF	48

Table 4.8: Service and arrival rates considered for the results on the effect of server disruption risk (Fig. 4.6).

Setting	**Service Rates**				**Arrival Rates**		
	μ_{11}	μ_{12}	μ_{22}	μ_{23}	λ_1	λ_2	λ_3
1	1.0	1.0	1.0	1.0	0.25	0.50	0.25
2	1.0	1.0	1.0	1.0	0.40	0.50	0.10
3	1.0	1.0	1.0	1.0	0.40	0.20	0.40
4	1.0	1.0	1.0	1.0	0.10	0.80	0.10
5	1.0	1.0	1.0	1.2	0.25	0.50	0.30
6	1.0	1.0	1.0	1.2	0.40	0.50	0.12
7	1.0	1.0	1.0	1.2	0.40	0.20	0.48
8	1.0	1.0	1.0	1.2	0.10	0.80	0.12
9	1.2	1.0	1.0	1.2	0.30	0.50	0.30
10	1.2	1.0	1.0	1.2	0.48	0.50	0.12
11	1.2	1.0	1.0	1.2	0.48	0.20	0.48
12	1.2	1.0	1.0	1.2	0.12	0.80	0.12

Table 4.9: Holding cost setting considered for the results on the effect of server disruption risk (Fig. 4.6).

Setting	**Holding Cost**		
	h_1	h_2	h_3
1	1	1	1
2	1	3	2
3	1	2	3
4	2	1	3

Table 4.10: Disruption and repair rate setting considered for the results on the effect of server disruption risk (Fig. 4.6).

Setting	**Disruption/Repair Rates**			
	r_1	r_2	θ_1	θ_2
1	0	0.5	0	0
2	0	0.5	0	0.1
3	0	0.5	0	0.2
4	0	0.5	0	0.3
5	0	0.5	0	0.4
6	0	0.5	0	0.5

CHAPTER 5

Supply Chain Disruption Risk Management: Newsvendor Analysis with Recourse

5.1 Introduction

In recent years, a variety of events have elevated to a strategic level concerns over the pernicious effects of supply chain disruptions. Consider the 2007 disruption in Boeing's 787 Dreamliner's supply chain. Advanced Integration Technology (AIT) fell months behind building parts needed to assemble the plane, thereby wreaking havoc upon Boeing's 787 inflexible supply chain. Boeing itself expected to take a cash hit of $2.5 billion in 2008 from paying penalties to airline customers and to keep its suppliers afloat in the wake of the serious cash flow disruption, according to [55]. Evidence showed that AIT had been facing serious production problems in 2006 (see, for instance, [55]). Thus, if Boeing had thoroughly monitored AIT in an effort to obtain *disruption risk information*, it might have made better ordering and contracting decisions in advance and thereby protected against such havoc. To enable firms to monitor their suppliers, some companies including Open Ratings have developed supply chain monitoring software that provides a firm with supplier visibility and actionable insights before a disruption occurs.

Consider also the disruption in the Toyota supply chain on Feb. 01, 1997. A fire at

the Aisin Seiki Co. destroyed most of the capacity to manufacture P-valves. Because of the Aisin's ability to produce parts at low cost, Toyota had come to rely on Aisin for this product ([129]). According to the Wall Street Journal, Toyota officials called different part makers to obtain P-valves, including Somic ([115]). Somic had the *flexibility* to free up machines and shift its production line to make P-valves. On Feb. 06, right on schedule, it delivered its first P-valves to Toyota ([115]). Considering the enormous financial impact of disruptions, it may be beneficial for many firms that procure raw materials from low price, high volume primary suppliers to reserve in advance some capacity from a *secondary flexible backup supplier* such as Somic to insure the supply stream against possible disruptions. This chapter gives insights into the potential benefits and risks.

Although disruptions are rare, their economic consequences can be massive. [64] investigate over 800 cases of disruptions in supply chains and conclude that firms suffering from supply chain disruptions experience about 30% lower stock returns than their matched benchmarks.

[92] have formulated a set of 10 principles derived from the supply chain risk literature, three of which are: (1) *diversification* in sourcing, (2) implementing *flexibility*, and (3) *information* sharing. These principles will be considered throughout this chapter from the perspective of a firm that procures materials from different sources.

We investigate two important strategic remedies to increase supply chain reliability and responsiveness: (1) contracting with a secondary flexible backup supplier (or similarly to establish an in-plant flexible resource) which is capable of adjusting its production mix to respond to the requests of a firm in the case of a disruption, and (2) to obtain and assess information about the disruption risk of primary suppliers. Point (1) increases the flexibility and responsiveness of supply chains in facing disruptions. In point (2), monitoring suppliers allows firms to anticipate potential disruptions and adopt better operational decisions.

Using a model with two products, we quantify the value of purchasing flexible

backup capacity through a *generalized capacity reservation* contract with a flexible backup supplier. In this contract, also known as an option contract, the buyer pays a fixed lump-sum payment to the supplier at the beginning of the contract in return for the delivery of any desired portion of the reserved fixed capacity at an additional proportional purchasing cost. (See [127] for the traditional case that has only the lump-sum cost.) If the buyer has enough on hand stock, less than the reserved capacity may be ordered to avoid additional holding and purchasing costs. Indeed, through this contract the buyer initially buys the option to order (any combination of the two products) up to a certain level from the supplier and later decides how to exercise this option by specifying the ordering quantity for each of the products. Moreover, based on the terms of the contract, the flexible backup supplier guarantees any amount of delivery up to the reserved capacity. Therefore, from the firm's point of view, this secondary supplier works as a reliable flexible backup (that can mitigate the risk of disruption in primary suppliers while reducing the cost of keeping excess inventory) with a limited flexible capacity in proportion to the up-front lump-sum investment.

Similar capacity reservation arrangements designed to provide the buyer with flexibility in the order quantity are observed in different high-tech industries such as semiconductors, consumer electronics, telecommunications, and pharmaceutical (where the demand for high-tech products is highly volatile and difficult to forecast) as well as in some automakers ([65]) and the textile and garment industry ([42]). Capacity reservation is also regarded as one of the counter-measures of the "bullwhip effect" ([95]).

In addition to providing insights into purchasing flexible backup capacity (through a capacity reservation contract), we investigate the value of the firm's monitoring of unreliable suppliers to obtain a more accurate perception of disruption risks. Our analyses quantify the financial impact of the misperception of the true reliabilities of suppliers.

To perform our analysis, we consider a two-stage setting where the firm is endowed with a recourse possibility to selectively utilize the capacity reserved with the secondary flexible backup supplier *after* monitoring the risk of primary suppliers. Using our two-stage setting, we also investigate the value of implementing flexibility in the backup system. The value of information on disruption risks is computed, optimal ordering decisions are identified, and the optimal capacity reservation level is quantified (including bounds).

The remainder of the chapter is organized as follows. We review the literature in next section, then in Section 6.3 we describe our model. Section 5.4 considers the problem in a two-stage setting with recourse. To gain insights into the value of recourse, in Section 5.5 we provide some benchmark analysis by considering the case where benchmark is not allowed. In Section 5.6, we use our framework to provide insights into the value of recourse, backup flexibility, and disruption risk information. Finally, we summarize our main findings in Section 5.7 and conclude.

5.2 Related Literature

This chapter considers *all-or-nothing* supply: when the supplier is up it delivers an order in full, while nothing can be supplied when it is down. By contrast, in models with *yield uncertainty* or *random yield*, the quantity received is a random fraction of the quantity ordered (see [41] for an comprehensive review of the literature).

The majority of supply-disruption papers focus on a single-supplier problem (see, for instance, [104], [19], [111], [112] , [57], [133], [107], [110], and [108]). [113] and [58] are among the supply disruption papers that consider more than one supplier. Both papers consider a firm that faces constant demand and sources from two identical-cost, infinite-capacity suppliers. [10] studies a finite-horizon, discrete-time, i.i.d. stochastic continuous demand model in which there are two zero lead time random-yield suppliers.

More recent related supply disruption papers with multiple suppliers include [14], [13], [145], [146], [32], [25], [43], [158], and [73]. [13] studies the effects of disruption risk in a supply chain where one retailer deals with competing risky suppliers who may default during their production lead-times.[14] uses a single-period model of a two-echelon supply chain with competing risky suppliers and a single firm and investigates how the supplier default risk and default co-dependence affect firm procurement and production decisions. [145] sheds light on some effective disruption risk mitigation mechanism by considering a single-product setting in which a firm can source from two suppliers, one that is unreliable and another that is completely reliable but more expensive and may possess volume flexibility. [146] investigates the value of a threat advisory system and develops multi-period models in which the firm has a single unreliable supplier, as well as models in which a second, perfectly reliable supplier is available. [32] considers a newsvendor that is served by multiple suppliers, where any given supplier is identified to be either perfectly reliable or unreliable. [25] considers the effect of decoupling delays (recurrent risk) and disruption risks in a model with two suppliers: an unreliable supplier and a perfectly reliable supplier that is under a capacity reservation contract. [43] considers supply diversification under general supply risk for a single product and a single demand season. [158] uses two-stage stochastic programming settings and investigates the value of two disruption mitigation mechanisms: dual sourcing and process improvement.

The effect of disruption information has also been studied under different settings in some recent papers (see, for example, [144], [163], and the references therein).

All the above-mentioned papers examine a single product setting. However, a single product setting does not allow us to capture the mix flexibility of a backup supplier such as Somic. To the best of our knowledge, this is the first work that develops two-product analyses to simultaneously consider the option of implementing mix flexibility in supply and the possibility of obtaining disruption risk information.

For reviews of flexibility, we refer readers to [128], [49], and [139]. In fact, the mix

flexibility of production operations has been studied in many papers including [46], [85], [154], [93], [53], [147], [82], and [79]. Our study contributes to this literature by considering the value of a flexible supplier/resource to compensate for the unreliability of dedicated suppliers. Similar contributions, but in the context of design of queueing systems with flexible resources, can be found in [124], and the references therein.

5.3 Model and Notation

Consider a centralized model of the contracting and ordering decisions of a firm (a manufacturer or retailer) in a two-echelon make-to-stock supply chain which produces/sells two types of products (namely 1 and 2) and has two dedicated suppliers, each capable of supplying units (components, final products, or raw materials) for one of the products. Denote the dedicated supplier of units of product j as supplier j (for $j = 1, 2$). The firm also has a flexible backup supplier (namely f) that can produce (under a capacity reservation contract) units for both products 1 and 2, the sum of which cannot exceed a reserved capacity, $\bar{Q}^f$. The capacity reservation contract explicitly allows the purchase of a flexible backup capacity, $\bar{Q}^f$. We typically use subscripts for products, superscripts for suppliers, and employ the following notation:

h_j : Holding cost per unit of product j; $(j = 1, 2)$

p_j : Lost sale penalty cost per unit of unmet demand of product j; $(j = 1, 2)$

r_j : Revenue per unit of product j sold; $(j = 1, 2)$

c^j : Per unit purchasing cost of product j from dedicated supplier j; $(j = 1, 2)$

c_j^f : Per unit purchasing cost of product j from flexible supplier; $(j = 1, 2)$

u^f : Per unit capacity reservation cost of the flexible backup supplier;

u_j^f : $(= u^f + c_j^f)$ unit cost of product j from the flexible backup supplier; $(j = 1, 2)$

q^j : Order quantity from dedicated supplier j; $(j = 1, 2)$

q_j^f : Order quantity from the flexible backup supplier for product j; $(j = 1, 2)$

$\bar{Q}^f$: Reserved capacity (in units) from the flexible backup supplier.

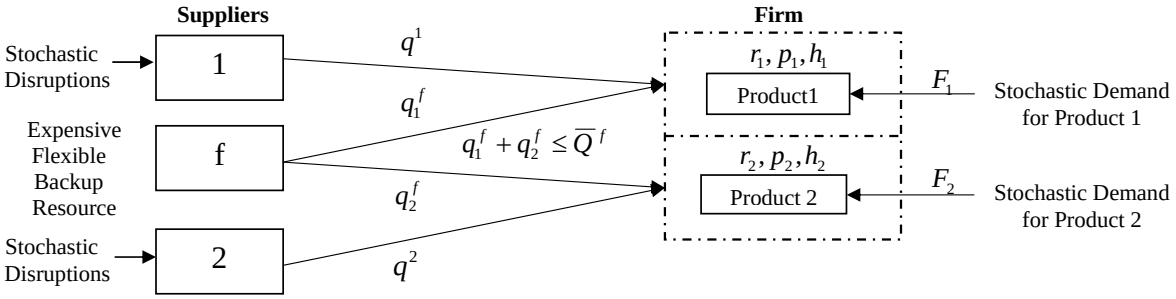


Figure 5.1: The two-echelon supply chain under consideration.

Fig. 1 depicts the two-echelon supply chain model. We assume the firm has a *price-only* contract with its unreliable dedicated suppliers, i.e., the firm pays c^j ($\leq r_j$) per unit delivered by dedicated (and unreliable) supplier j. Moreover, it has a generalized capacity reservation contract with its flexible backup supplier, i.e., the firm chooses the capacity level of $\bar{Q}^f$ units and pays the flexible backup supplier a cost of $u^f \times \bar{Q}^f$ at time 0 to reserve its capacity. The flexible backup supplier, in return, agrees to deliver any order of products 1 and 2 (q_1^f and q_2^f) subject to $q_1^f + q_2^f \leq \bar{Q}^f$, and the firm pays the purchasing cost of $c_1^f\, q_1^f + c_2^f\, q_2^f$. We allow $u_j^f = u^f + c_j^f$ to be greater or less than r_j. However, to avoid the trivial case where single sourcing from the flexible backup supplier is always optimal, we shall assume $u_j^f = u^f + c_j^f > c^j$. This corresponds to practice, because the contract can be regarded as an investment in the flexible supplier's technology which is not cheaper than inflexible ones. We also note that the flexible supplier in our framework could be an in-plant flexible resource that requires an upfront investment of u^f per unit of capacity.

For j ($j = 1, 2$), let $\mathcal{L}_j(x) = h_j[x]^+ + p_j[-x]^+$ and

$$L_j(x) = E_{D_j}[\mathcal{L}_j(x - D_j)] = h_j \int_0^x (x - \xi) dF_j(\xi) + p_j \int_x^\infty (\xi - x) dF_j(\xi), \qquad (5.1)$$

where $[x]^+ = \max\{0, x\}$, and $F_j(\cdot)$ is the cumulative distribution function (CDF) of the demand for product j (random variable D_j). We assume $F_j(\cdot)$ (for both $j = 1, 2$) is differentiable and has $F_j(x) = 0$ for all $x \leq 0$ and $F_j(x) > 0$ for all $x > 0$.

Denoting the survival function (i.e., the complementary CDF) of demand of product j by $\bar{F}_j(\cdot) = 1 - F_j(\cdot)$, we define the inventory cost function $G_j(\cdot)$ as:

$$G_j(x) = L_j(x) - r_j\, E[\min(D_j\,,\, x)] = L_j(x) - r_j \int_0^x \bar{F}_j(\xi) d\xi, \tag{5.2}$$

where the last expression can be obtained via integration by parts. In these definitions, $L_j(\cdot)$ is the expected total inventory cost of product j (holding plus shortage) and $G_j(\cdot)$ is the expected cost minus the expected revenue obtained from product j.

We note that a firm may not accurately perceive the disruption risk of its unreliable suppliers. However, we can model the firm's best estimate of the reliability of supplier j (i.e., the probability that supplier j is up) by θ^j. We define $\Theta = (\theta^1, \theta^2)$ as the vector of *perceived* reliabilities and $\Pi = (1 - \pi_0^1\,, 1 - \pi_0^2)$ as the vector of *true* reliabilities. Also, we let $\Upsilon = (\epsilon^1, \epsilon^2)$ denote the firm's error in estimating the true reliability vector where $\Theta = \Pi + \Upsilon$ (i.e., $\theta^j = 1 - \pi_0^j + \epsilon^j$ for both $j = 1, 2$).

5.4 Analyses with Recourse (Two-Stage Setting)

To generate insights into effective disruption mitigation mechanisms for firms, we start by analyzing the case where the firm can first monitor the "availability" (i.e., up or down) state of its primary suppliers and then utilize a recourse option of ordering from the secondary flexible backup supplier. Note that the availability state may be construed as the ability of a supplier to deliver the requested products within a required time. In other words, one can also think of up (down) state in terms of on time (late) delivery.

The sequence of events in this two-stage scenario is as follows. The firm first decides to reserve a capacity of $\bar{Q}^f$ units from the secondary flexible backup supplier and pays $u^f \times \bar{Q}^f$ to do so (Stage 1). Then the firm observes the state of its primary suppliers, purchases from the available ones, and uses the flexible backup supplier subject to the reserved capacity (Stage 2). Then demands are realized and inventory

costs (inventory shortage or holding minus the sales revenue) accrue. Observe that in Stage 1, the firm can insure the supply stream against possible disruptions through investment in a flexible backup capacity. It can then monitor the suppliers and use this information to make better ordering decisions. Indeed, in Stage 1, the firm can purchase the (recourse) option to benefit from flexible backup capacity proportional to its investment. In stage 2, after observing the disruption states, the firm can exercise this option at a cost of c_j^f per mix of type j. After analyzing this sequence of events, in Section 5.4.3, we will introduce an alternative sequence of events to allow consideration of offshore unreliable suppliers.

Using the above-mentioned framework we seek to answer the following questions.

Question 1 *If the reserved capacity in Stage 1 is limited, how would the firm* distribute *the available flexible backup capacity among its products based on the obtained information?*

Question 2 *How much would a firm invest in the flexible supplier as a backup for possible disruptions? Does obtaining a recourse option result in a reduction in such an investment?*

Also, comparing the scenario with recourse to the benchmark setting described in Section 5.5 (i.e., when the firm cannot observe the states of unreliable suppliers before ordering), we can ask the following question.

Question 3 *How beneficial is having a recourse option for firms, and can it be regarded as a strong risk mitigation mechanism?*

Moreover, it is interesting to compare a scenario with two dedicated backup suppliers (one for each product) with the one with a single but flexible backup supplier to answer the following question.

Question 4 *How beneficial is the flexibility of a backup supplier, and for what firms should implementing flexibility (in the backup system) be more attractive?*

Finally, similar to our no-recourse setting, we want to investigate the value of disruption risk information (that can reduce the risk belief errors) and address the following question.

Question 5 *How valuable is obtaining disruption risk information (under a recourse option) for firms, and can it be regarded as a strong risk mitigation mechanism?*

To answer such questions, let $C_{U,U}(\bar{Q}^f)$, $C_{U,D}(\bar{Q}^f)$, $C_{D,U}(\bar{Q}^f)$, and $C_{D,D}(\bar{Q}^f)$ (U: Up, D: Down) denote the minimum expected cost of the firm in the second stage, if both suppliers are up, only the first supplier is up, only the second supplier is up, and when none of them are up, respectively. These costs can be computed as follows.

$$\begin{aligned} C_{U,U}(\bar{Q}^f) &= \min_{q^1,q^2,q_1^f,q_2^f \geq 0 \text{ s.t. } q_1^f+q_2^f \leq \bar{Q}^f} \sum_{j=1}^{2} c_j^f\, q_j^f + c^1\, q^1 + c^2\, q^2 + G_1(q_1+q_1^f) \\ &\qquad +G_2(q_2+q_2^f) \qquad (5.3) \\ C_{U,D}(\bar{Q}^f) &= \min_{q^1,q_1^f,q_2^f \geq 0 \text{ s.t. } q_1^f+q_2^f \leq \bar{Q}^f} \sum_{j=1}^{2} c_j^f\, q_j^f + c^1\, q^1 + G_1(q_1+q_1^f) + G_2(q_2^f) \qquad (5.4) \\ C_{D,U}(\bar{Q}^f) &= \min_{q^2,q_1^f,q_2^f \geq 0 \text{ s.t. } q_1^f+q_2^f \leq \bar{Q}^f} \sum_{j=1}^{2} c_j^f\, q_j^f + c^2\, q^2 + G_1(q_1^f) + G_2(q_2+q_2^f) \qquad (5.5) \\ C_{D,D}(\bar{Q}^f) &= \min_{q_1^f,q_2^f \geq 0 \text{ s.t. } q_1^f+q_2^f \leq \bar{Q}^f} \sum_{j=1}^{2} c_j^f\, q_j^f + G_1(q_1^f) + G_2(q_2^f). \qquad (5.6) \end{aligned}$$

Now, if $C_{Stage\,2}(\bar{Q}^f)$ represents the optimal expected cost of Stage 2 as is perceived by the firm at the beginning of Stage 1, we have: $C_{Stage\,2}(\bar{Q}^f) =$

$$\theta^1\,\theta^2\,C_{U,U}(\bar{Q}^f) + \theta^1\,(1-\theta^2\,)C_{U,D}(\bar{Q}^f) + (1-\theta^1)\,\theta^2\,C_{D,U}(\bar{Q}^f) + (1-\theta^1)(1-\theta^2)\,C_{D,D}(\bar{Q}^f). \qquad (5.7)$$

Then, the firm can determine the optimal capacity reservation level ($\bar{Q}^{f*}$) at the beginning of Stage 1 by solving the following program:

$$\min_{\bar{Q}^f \geq 0} \quad u^f\,\bar{Q}^f + C_{Stage\,2}(\bar{Q}^f). \qquad (5.8)$$

To determine the behavior of the firm, we must first optimize non-linear programs (5.3)-(5.6) of Stage 2 for a given capacity reservation level; thereafter, the optimal ordering policy can be used to derive the optimal contracting level as is perceived by the firm by solving program (5.8). This optimal capacity reservation level (a strategic decision) then determines the tactical ordering behavior in each case (i.e. the minimizers of (5.3)-(5.6)). To solve programs (5.3)-(5.6), we note that the objective functions are all jointly convex in their variables (see Appendix C for the proof of Lemma 7 and of our other results), and the constraints are linear. Hence, the KKT conditions are sufficient and necessary to characterize the optimal solutions. To gain some insights, we first start by considering a single-product setting as a special case, and next solve (5.3)-(5.6) to derive the optimal policy of the firm under the original two-product setting.

5.4.1 Single–Product Special Case

Consider a single-product version of the problem discussed in the previous section. Since there is only one product and one unreliable supplier, we suppress both the product index, j, and the index of unreliable suppliers. However, we continue to use index f to denote the backup supplier. To characterize the firm's optimal capacity reservation level from the backup supplier as well as its optimal ordering policy, we need to first solve the problem in Stage 2 (i.e., derive the firm's optimal ordering policy) for any level of capacity reserved with the backup supplier in Stage 1. Subsequently, we can use the obtained results to solve Stage 1 to find the optimal capacity reservation level.

Proposition 8 (Single-Product) *Given that the firm reserves $\bar{Q}^f$ units of capacity from the backup supplier in Stage 1, the optimal ordering policy of Stage 2 is as follows.*
(1) If $c^f > c$ and the unreliable supplier is observed to be up, then $q^{f*} = 0$, and $q^* = F^{-1}(\frac{p+r-c}{p+r+h})$.

(2) If $c^f > c$ and the unreliable supplier is observed to be down, then

$q^{f*} = \min\{F^{-1}(\frac{p+r-c^f}{p+r+h}), \bar{Q}^f\}$.

(3) If $c^f \leq c$ and the unreliable supplier is observed to be up, then

(i) If $\bar{Q}^f \in \left[F^{-1}(\frac{p+r-c^f}{p+r+h}), +\infty \right)$ *then* $q^{f*} = F^{-1}(\frac{p+r-c^f}{p+r+h})$ *and* $q^* = 0$.

(ii) If $\bar{Q}^f \in \left[F^{-1}(\frac{p+r-c}{p+r+h}), F^{-1}(\frac{p+r-c^f}{p+r+h}) \right)$ *then* $q^{f*} = \bar{Q}^f$ *and* $q^* = 0$.

(iii) If $\bar{Q}^f \in \left[0, F^{-1}(\frac{p+r-c}{p+r+h}) \right)$ *then* $q^{f*} = \bar{Q}^f$ *and* $q^* = F^{-1}(\frac{p+r-c}{p+r+h}) - \bar{Q}^f$.

(4) If $c^f \leq c$ and the unreliable supplier is observed to be down, then

(i) If $\bar{Q}^f \in \left[F^{-1}(\frac{p+r-c^f}{p+r+h}), +\infty \right)$ *then* $q^{f*} = F^{-1}(\frac{p+r-c^f}{p+r+h})$.

(ii) If $\bar{Q}^f \in \left[0, F^{-1}(\frac{p+r-c^f}{p+r+h}) \right)$ *then* $q^{f*} = \bar{Q}^f$.

Parts (1) and (2) of Proposition 8 describe that when exercising the option from the backup supplier is more expensive than purchasing from the unreliable supplier, the firm orders only from a single source: it only uses the unreliable supplier if it is up, and only uses the backup supplier (as much as required or the reserved capacity allows) otherwise. As parts (3) and (4) of Proposition 8 show, when exercising the option from the backup supplier is cheaper than purchasing from the unreliable supplier, if the unreliable supplier is observed to be up, then the firm either orders nothing from the unreliable supplier or exhausts the capacity reservation option (unless the reserved capacity is sufficient) and only orders the rest of its requirement from the unreliable supplier. On the other hand, when the unreliable supplier is observed to be down, the firm exhausts the capacity reservation option (unless the reserved capacity is sufficient). In fact, the firm's problem in a single-product setting is much easier than the two-product setting, since it is not facing the complex problem of rationing the available limited *flexible* backup capacity between the two products in an effective way. As we will show, in the two-product setting, even when only one of the suppliers is down, the firm may need to ration the backup flexible capacity between the two products (see Theorem 8 part (ii)).

Remark. A special case of the single-product version of our model presented in this section is the case where the cost of reserving capacity in Stage 1 (u^f) is negligible. In this very special case where (1) the backup supplier is not endowed with mix flexibility (since it is a single-product setting), and (2) backup capacity is unlimited (i.e., $\bar{Q}^f$ is large enough) and does not need to be reserved in advance, the role of the backup supplier can be construed via the second opportunity quick response models studied thoroughly in papers such as Fisher and Raman (1996) (the Sport Obermeyer case), Eppen and Iyer (1997), Milner and Kouvelis (2002), and Li et. al (2009). The main difference between our work and such models, even under the special case of a single-product setting, is that the firm must optimize the amount of backup capacity to be reserved in advance (i.e., in an anticipation of potential future disruptions) subject to a reservation cost. Indeed, the firm in our model can insure the supply stream by buying backup capacity in advance. Note that this investment greatly affects the ordering ability of the firm in the second stage. Furthermore, another main feature of our model that differentiates it from such studies is the mix flexibility of the backup flexible supplier that we will consider in the sequel. In the presence of a backup capacity that is (a) limited and (b) flexible, we provide insights into the question of how to effectively benefit from a recourse option and *ration* the limited backup capacity between products to compensate for the disruption risk of primary suppliers. Comparing our setting with a setting where the backup capacity is not flexible (i.e., a setting with two independent products), we will see that the mix flexibility in the backup system provides a significant advantage for the firm in the presence of unreliable suppliers. Another distinct and novel objective of our research is to provide insights into the value of obtaining disruption risk information as we will discuss in Sections 5.5.2 and 5.6.3.

Having the ordering policy in hand, we can now use Proposition 8 to determine the optimal capacity reservation level of Stage 1.

Proposition 9 (Capacity Reservation Level) *The optimal capacity reservation*

level (as perceived by the firm) can be characterized as:

(1) If $c^f > c$*, then*

$$\bar{Q}^{f*} = F^{-1}([\frac{p + r - \frac{u^f + (1-\theta)c^f}{1-\theta}}{p + r + h}]^+).$$

(2) If $c^f \leq c$*, then*

$$\bar{Q}^{f*} = F^{-1}([\frac{p + r - \frac{u^f + c^f - \theta c}{1-\theta}}{p + r + h}]^+).$$

Part (1) and (2) of the above proposition can be combined and summarized as follows. Let $\bar{c} = \max\{c, c^f\}$ and $\hat{c} = (u^f + c^f - \theta\bar{c})/(1-\theta)$. Then $\bar{Q}^{f*} = F^{-1}([\frac{p+r-\hat{c}}{p+r+h}]^+)$, which can be construed as a single-source (and single-stage) traditional newsvendor setting with a perfectly reliable source and a purchasing cost $\hat{c}$. However, note that this new purchasing cost, $\hat{c}$, is affected by the reliability perception of the firm, θ, as well as the backup capacity reservation costs, u^f and c^f. For instance, the firm will not reserve any backup capacity in Stage 1 if its reliability perception of the unreliable supplier is greater than a threshold (that depends on the inventory, purchasing, and capacity reservation costs), even though it might be truly beneficial to reserve some backup capacity.

5.4.2 Two–Product Case

We now consider our original two-product setting to provide insights into Questions 1-5. The following three theorems solve programs (5.3)-(5.6) to define the optimal ordering policy of the firm and provide insight into Question 1. For brevity, we only consider the most interesting situation where $c_j^f \leq c^j$ throughout this section, but other situations can be analyzed in a similar way (see Appendix D for the case where $c_j^f > c^j$).

Theorem 7 (Both Suppliers Up) *Let* $k = \text{argmax}\{c^j - c_j^f : j = 1, 2\}$ *denote the product with the higher difference in purchasing cost, and* $l = 3 - k$ *be the other product. If both suppliers are observed to be up, then the following cases fully characterize the optimal ordering policy of the firm, given a reserved flexible backup capacity of*

$\bar{Q}^f$:

(i) *If* $\bar{Q}^f \in \left[\sum_{j=k,l} F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j}), +\infty \right)$ *then* $q_j^{f*} = F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j})$ *and* $q^{j*} = 0$ $(j = k, l)$.

(ii) *If* $\bar{Q}^f \in \left[\sum_{j=k,l} F_j^{-1}(\frac{p_j+r_j-c_j^f-(c^l-c_l^f)}{p_j+r_j+h_j}), \sum_{j=k,l} F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j}) \right)$ *then* $q_j^{f*} = F_j^{-1}(\frac{p_j+r_j-c_j^f-t}{p_j+r_j+h_j})$ *and* $q^{j*} = 0$ $(j = k, l)$, *where* $t \in (0, c^l - c_l^f]$ *is the solution to* $\sum_{j=k,l} F_j^{-1}(\frac{p_j+r_j-c_j^f-t}{p_j+r_j+h_j}) = \bar{Q}^f$.

(iii) *If* $\bar{Q}^f \in \left[F_k^{-1}(\frac{p_k+r_k-c_k^f-(c^l-c_l^f)}{p_k+r_k+h_k}), \sum_{j=k,l} F_j^{-1}(\frac{p_j+r_j-c_j^f-(c^l-c_l^f)}{p_j+r_j+h_j}) \right)$ *then* $q_k^{f*} = F_k^{-1}(\frac{p_k+r_k-c_k^f-(c^l-c_l^f)}{p_k+r_k+h_k})$, $q_l^{f*} = \bar{Q}^f - q_k^{f*}$, $q^{k*} = 0$, *and* $q^{l*} = F_l^{-1}(\frac{p_l+r_l-c^l}{p_l+r_l+h_l}) - q_l^{f*}$.

(iv) *If* $\bar{Q}^f \in \left[F_k^{-1}(\frac{p_k+r_k-c^k}{p_k+r_k+h_k}), F_k^{-1}(\frac{p_k+r_k-c_k^f-(c^l-c_l^f)}{p_k+r_k+h_k}) \right)$ *then* $q_k^{f*} = \bar{Q}^f$, $q_l^{f*} = 0$, $q^{k*} = 0$, *and* $q^{l*} = F_l^{-1}(\frac{p_l+r_l-c^l}{p_l+r_l+h_l})$.

(v) *If* $\bar{Q}^f \in \left[0, F_k^{-1}(\frac{p_k+r_k-c^k}{p_k+r_k+h_k}) \right)$ *then* $q_k^{f*} = \bar{Q}^f$, $q_l^{f*} = 0$, $q^{k*} = F_k^{-1}(\frac{p_k+r_k-c^k}{p_k+r_k+h_k}) - \bar{Q}^f$, *and* $q^{l*} = F_l^{-1}(\frac{p_l+r_l-c^l}{p_l+r_l+h_l})$.

Theorem 7 part (i) shows that if the firm has already reserved more than enough capacity (in Stage 1), it would not order anything from the primary suppliers (in Stage 2) and will only use the available flexible capacity, ordering the optimal level for each of the products. The rest of the reserved capacity is wasted to avoid paying extra holding or purchasing costs. Even if the reserved capacity is below the level identified in part (i), but it is in the range described by part (ii), the firm will only use the reserved capacity. However, in this case, it will appropriately *ration* $\bar{Q}^f$ between the products. Indeed, t can be viewed as a fictitious additional ordering cost that is applied to optimally ration the available flexible capacity. Part (iii) states that if the available capacity is enough to fulfill a prescribed amount of product k, but not enough for both of the products, the firm should use the flexible capacity to satisfy all of the optimal ordering amount of product k. Hence, it would not order anything from primary supplier k, and use the rest of the reserved capacity as well as primary

supplier l to satisfy the requirement product l. In the case that the reserved capacity is not enough to meet the prescribed level for any of the products, but still is relatively large, part (iv) shows that the firm should set aside all the reserved flexible capacity for product k (and not order from dedicated supplier k). Hence, in this case, it is optimal to only use primary supplier l to optimize the service level of product l. If the reserved flexible capacity is very low, as is presented in part (v), it is optimal to use all $\bar{Q}^f$ units of the limited flexible capacity for the "expensive" product (i.e., product k), and also procure the rest of requirements of this product from its primary supplier. Moreover, similar to case (iv), product l is sourced only through its primary supplier. We now treat the case when exactly one of the suppliers is observed to be disrupted.

Theorem 8 (One Supplier Up) *Let $m \in \{1,2\}$ denote the dedicated supplier that is observed to be up, and $n = 3 - m$ be the disrupted supplier. Then the following cases fully characterize the optimal ordering policy of the firm, given a reserved flexible backup capacity of $\bar{Q}^f$:*

(i) If $\bar{Q}^f \in \left[\sum_{j=n,m} F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j}) , +\infty \right)$ then $q_j^{f*} = F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j})$ $(j = n, m)$ and $q^{m*} = 0$.

(ii) If $\bar{Q}^f \in \left[\sum_{j=n,m} F_j^{-1}([\frac{p_j+r_j-c_j^f-(c^m-c_m^f)}{p_j+r_j+h_j}]^+) , \sum_{j=n,m} F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j}) \right)$ then $q_j^{f*} = F_j^{-1}([\frac{p_j+r_j-c_j^f-t}{p_j+r_j+h_j}]^+)$ $(j = n, m)$ and $q^{m*} = 0$, where $t \in (0, c^m - c_m^f]$ is the solution to $\sum_{j=n,m} F_j^{-1}([\frac{p_j+r_j-c_j^f-t}{p_j+r_j+h_j}]^+) = \bar{Q}^f$.

(iii) If $\bar{Q}^f \in \left[F_n^{-1}([\frac{p_n+r_n-c_n^f-(c^m-c_m^f)}{p_n+r_n+h_n}]^+) , \sum_{j=n,m} F_j^{-1}([\frac{p_j+r_j-c_j^f-(c^m-c_m^f)}{p_j+r_j+h_j}]^+) \right)$ then $q_n^{f*} = F_n^{-1}([\frac{p_n+r_n-c_n^f-(c^m-c_m^f)}{p_n+r_n+h_n}]^+)$, $q_m^{f*} = \bar{Q}^f - q_n^{f*}$, and $q^{m*} = F_m^{-1}(\frac{p_m+r_m-c^m}{p_m+r_m+h_m}) - q_m^{f*}$.

(iv) If $\bar{Q}^f \in \left[0 , F_n^{-1}([\frac{p_n+r_n-c_n^f-(c^m-c_m^f)}{p_n+r_n+h_n}]^+) \right)$ then $q_n^{f*} = \bar{Q}^f$, $q_m^{f*} = 0$, and $q^{m*} = F_m^{-1}(\frac{p_m+r_m-c^m}{p_m+r_m+h_m})$.

It is noteworthy that Theorem 8 can be interpreted via Theorem 7. In fact, by considering the disrupted supplier as a supplier with a sufficiently large purchasing cost, Theorem 8 coincides with Theorem 7. However, in Theorem 8 the operator $[\cdot]^+$ is used (whenever necessary) to assure that the ratios are within appropriate domain of functions $F_j^{-1}(\cdot)$. This is redundant in Theorem 7 because of the assumption $r_j \geq c^j = c_j^f + (c^j - c_j^f)$ (for both j=k,l), which also implies $r_k \geq c_k^f + (c^l - c_l^f)$ since $c^k - c_k^f \geq c^l - c_l^f$. To conclude, the following theorem treats the case when both primary suppliers are disrupted, and similarly can be interpreted via Theorem 7 with sufficiently large c^j $(j = 1, 2)$.

Theorem 9 (Both Suppliers Disrupted) *The following cases fully characterize the optimal ordering policy of the firm when both primary suppliers are observed to be disrupted, given a reserved flexible backup capacity of* $\bar{Q}^f$*:*

(i) If $\bar{Q}^f \in \left[\sum_{j=1,2} F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j}) \, , +\infty \right)$ *then* $q_j^{f*} = F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j})$ *$(j = 1, 2)$.*

(ii) If $\bar{Q}^f \in \left[0 \, , \sum_{j=1,2} F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j}) \right)$ *then* $q_j^{f*} = F_j^{-1}([\frac{p_j+r_j-c_j^f-t}{p_j+r_j+h_j}]^+)$ *$(j = 1, 2)$, where* $t \in (\, 0 \, , \max_j\{p_j + r_j - c_j^f\}\,]$ *is the solution to* $\sum_{j=1,2} F_j^{-1}([\frac{p_j+r_j-c_j^f-t}{p_j+r_j+h_j}]^+) = \bar{Q}^f$.

Theorems 7, 8, and 9 solve programs (5.3)-(5.6). Note that the separability of the ordering policy between products as reflected in Theorem 13 no longer applies. However, to answer Question 2, one can use Theorems 7, 8, and 9 to solve program (5.8) and obtain the optimal capacity reservation level (for any parameter setting and any demands distributions $F_1(\cdot), F_2(\cdot)$). For brevity, and to further analytically characterize the optimal contracting level, however, we only consider the case where the following mild assumption holds. For simplicity of presentation, it is also convenient to fix the labeling of the products such that product 2 is the product with higher difference in purchasing cost (i.e., $\text{argmax}\{c^j - c_j^f : j = 1, 2\} = 2$).

Assumption 1 *Demand distributions* $F_1(\cdot), F_2(\cdot)$ *are such that:*

(a) $0 \leq F_1^{-1}(\frac{p_1+r_1-c_1^f-(c^2-c_2^f)}{p_1+r_1+h_1}) \leq F_2^{-1}(\frac{p_2+r_2-c^2}{p_2+r_2+h_2})$, *and*
(b) $F_2^{-1}(\frac{p_2+r_2-c_2^f-(c^1-c_1^f)}{p_2+r_2+h_2}) \leq \sum_{j=1,2} F_j^{-1}(\frac{p_j+r_j-c_j^f-(c^2-c_2^f)}{p_j+r_j+h_j})$.

Notice that the above assumption is not critical, since (1) if it does not hold, analysis will follow similar lines, and (2) it holds for most settings where products are not extremely different in their procurement and inventory costs as well as their demand distributions. For instance, a completely symmetric scenario where the parameters are product-independent satisfies conditions (a) and (b).

The following proposition further helps to characterize the optimal capacity reservation level with the flexible backup supplier. It states that $\bar{Q}^{f*}$ can only take one of six possible values. Therefore, it can be found by comparing the cost of these six options.

Proposition 10 (Capacity Reservation Level) *Under Assumption 1,* $\bar{Q}^{f*} \in \{\bar{Q}_k^f : k = 1, 2, ..., 6\}$, *where*

$$\begin{aligned}
\bar{Q}_1^f &= \underset{\bar{Q}^f \in I_1}{\operatorname{argmin}} \big(u^f - \theta^1\theta^2(c^2 - c_2^f) + \theta^1\bar{\theta}^2 c_2^f + \bar{\theta}^1\theta^2 c_1^f\big)\bar{Q}^f \\
&\qquad\qquad + \bar{\theta}^1\theta^2 G_1(\bar{Q}^f) + \theta^1\bar{\theta}^2 G_2(\bar{Q}^f) + \bar{\theta}^1\bar{\theta}^2\Gamma(\bar{Q}^f), \\
\bar{Q}_2^f &= \underset{\bar{Q}^f \in I_2}{\operatorname{argmin}} \big(u^f - \theta^2(c^2 - c_2^f) + \theta^1\bar{\theta}^2 c_2^f\big)\bar{Q}^f + \theta^1\bar{\theta}^2 G_2(\bar{Q}^f) + \bar{\theta}^1\bar{\theta}^2\Gamma(\bar{Q}^f), \\
\bar{Q}_3^f &= \underset{\bar{Q}^f \in I_3}{\operatorname{argmin}} \big(u^f - \bar{\theta}^1\theta^2(c^2 - c_2^f) + \theta^1 c_2^f\big)\bar{Q}^f + \theta^1 G_2(\bar{Q}^f) + \bar{\theta}^1\bar{\theta}^2\Gamma(\bar{Q}^f), \\
\bar{Q}_4^f &= \underset{\bar{Q}^f \in I_4}{\operatorname{argmin}} \big(u^f - \theta^1(c^1 - c_1^f) - \bar{\theta}^1\theta^2(c^2 - c_2^f)\big)\bar{Q}^f + \bar{\theta}^1\bar{\theta}^2\Gamma(\bar{Q}^f), \\
\bar{Q}_5^f &= \underset{\bar{Q}^f \in I_5}{\operatorname{argmin}} \big(u^f - \theta^1(c^1 - c_1^f)\big)\bar{Q}^f + \bar{\theta}^1\Gamma(\bar{Q}^f), \\
\bar{Q}_6^f &= \underset{\bar{Q}^f \in I_6}{\operatorname{argmin}}\, u^f\bar{Q}^f + \Gamma(\bar{Q}^f),
\end{aligned}$$

with $\bar{\theta}^j = 1 - \theta^j$, $\Gamma(\bar{Q}^f) = \sum_{j=1,2}\left[c_j^f F_j^{-1}(\frac{p_j+r_j-c_j^f-t_{\bar{Q}^f}}{p_j+r_j+h_j}) + G_j(F_j^{-1}(\frac{p_j+r_j-c_j^f-t_{\bar{Q}^f}}{p_j+r_j+h_j}))\right]$ *($t_{\bar{Q}^f}$ is the solution to* $\sum_{j=1,2} F_j^{-1}(\frac{p_j+r_j-c_j^f-t_{\bar{Q}^f}}{p_j+r_j+h_j}) = \bar{Q}^f$*), and*

$$I_1 = \big[0, F_1^{-1}(\frac{p_1 + r_1 - c_1^f - (c^2 - c_2^f)}{p_1 + r_1 + h_1})\big],$$
$$I_2 = \big[F_1^{-1}(\frac{p_1 + r_1 - c_1^f - (c^2 - c_2^f)}{p_1 + r_1 + h_1}), F_2^{-1}(\frac{p_2 + r_2 - c^2}{p_2 + r_2 + h_2})\big]$$
$$I_3 = \big[F_2^{-1}(\frac{p_2 + r_2 - c^2}{p_2 + r_2 + h_2}), F_2^{-1}(\frac{p_2 + r_2 - c_2^f - (c^1 - c_1^f)}{p_2 + r_2 + h_2})\big],$$
$$I_4 = \big[F_2^{-1}(\frac{p_2 + r_2 - c_2^f - (c^1 - c_1^f)}{p_2 + r_2 + h_2}), \sum_{j=1,2} F_j^{-1}(\frac{p_j + r_j - c_j^f - (c^2 - c_2^f)}{p_j + r_j + h_j})\big]$$
$$I_5 = \big[\sum_{j=1,2} F_j^{-1}(\frac{p_j + r_j - c_j^f - (c^2 - c_2^f)}{p_j + r_j + h_j}), \sum_{j=1,2} F_j^{-1}(\frac{p_j + r_j - c_j^f - (c^1 - c_1^f)}{p_j + r_j + h_j})\big]\big],$$
$$I_6 = \big[\sum_{j=1,2} F_j^{-1}(\frac{p_j + r_j - c_j^f - (c^1 - c_1^f)}{p_j + r_j + h_j}), \sum_{j=1,2} F_j^{-1}(\frac{p_j + r_j - c_j^f}{p_j + r_j + h_j})\big].$$

We now present a corollary of the above proposition which provides two upper bounds on the optimal capacity reservation (or investment) level with the flexible backup supplier. The first upper bound is the sum of optimal orders to two separate, reliable, and dedicated suppliers with purchasing costs c_j^f $(j = 1, 2)$. This bound shows that the flexibility of the backup supplier offers competitive advantage to the firm. The second bound is the optimal capacity reservation level without the recourse option (see Section 5.5 for analysis without the recourse option). Indeed, the second part of the following corollary provides more insights to Question 2 by presenting a condition under which obtaining a recourse option will result in a non-strict reduction in the capacity reserved (or the investment level) with the flexible supplier.

Corollary 2 (Bounds) *The optimal capacity reservation level with the flexible backup supplier (with recourse) is bounded above by (i)* $\sum_{j=1,2} F_j^{-1}(\frac{p_j + r_j - c_j^f}{p_j + r_j + h_j})$*, and (ii) the optimal capacity reservation level without recourse, if disruption risks are such that* $\theta^j\,(c^j - c_j^f) \geq u^f$ *for* $j = 1, 2$.

To provide more insights, we now characterize the optimal backup capacity reservation level under a symmetric scenario where the two products have similar characteristics (demand distributions, procurement and inventory costs, and revenues – but not

necessarily supplier reliabilities). Such a symmetric scenario allow us to completely characterize the optimal capacity reservation level and gain some sharp insights. To this end, the following theorem considers a symmetric scenario and characterizes both cases where the flexible secondary supplier is more expensive or cheaper than the primary ones.

Theorem 10 (Symmetric Scenario) *Consider a symmetric scenario where all parameters except (perhaps) the supplier reliabilities are product-independent. Further, assume that product demands follow a uniform distribution between 0 and d (with* $d > \frac{2\,(p+r-u^f-c^f)}{p+r+h}$ *to ensure full linearity of the demand CDFs in the working range). Then, if* $c - c^f \geq u^f$*,we have:*

$$\bar{Q}^{f*} = \frac{2d\,(p+r-u^f-c^f)}{p+r+h}. \tag{5.9}$$

However, if $c-c^f < u^f$*, consider the following conditions on the reliability of suppliers: Condition 1 (C1):* $\bar{\theta}^1\,\bar{\theta}^2 \geq \frac{2\left(u^f-(c-c^f)\right)}{p+r-c}$*, and Condition 2 (C2):* $\theta^1\,\theta^2 \leq 1 - \frac{u^f-(c-c^f)}{p+r-c}$*. Then*

(i) when both C1 and C2 do not hold: $\bar{Q}^{f*} = 0$*,*

(ii) when C1 does not hold but C2 holds:

$$\bar{Q}^{f*} = \frac{d\left((1-\theta^1\theta^2)(p+r)+\theta^1\theta^2 c-u^f-c^f\right)}{(\bar{\theta}^1\theta^2+\theta^1\bar{\theta}^2+\frac{\bar{\theta}^1\bar{\theta}^2}{2})(p+r+h)}, \tag{5.10}$$

(iii) when C1 holds:

$$\bar{Q}^{f*} = \frac{2d\left(p+r-c-\frac{u^f-(c-c^f)}{\theta^1\,\theta^2}\right)}{p+r+h}. \tag{5.11}$$

The above theorem describes that, when the procurement cost from the flexible supplier $(u^f + c^f)$ is sufficiently cheap, the optimal capacity reservation level is independent of disruption risks. Furthermore, similar to the case without recourse (see

Theorem 13), there is a separation of the joint capacity reservation in this case; the optimal capacity reserved presented in (5.9) is the sum of the orders to two independent newsvendors with a procurement cost of $u^f + c^f$. However, when the procurement cost from the flexible supplier is not cheap, the separation phenomenon no longer exists. For instance, (5.11) shows that the optimal capacity reservation level with a recourse option is a function of $\bar{\theta}^1 \times \bar{\theta}^2$. However, as we will see in the case without recourse (see Theorem 13), the optimal capacity reservation level is the sum of two independent terms: one a function of $\bar{\theta}^1$ and the other a function of $\bar{\theta}^2$.

5.4.3 Recourse Analysis with Offshore Unreliable Suppliers

In some situations, a firm may not be able to monitor its unreliable suppliers before placing the orders. For instance, unlike our motivating examples discussed in the Introduction, there might be geographical or other barriers between such suppliers and the firm which prevent the firm from effectively monitoring its unreliable suppliers. Thus, we now assume that the firm first places the orders with unreliable suppliers, observes the delivered quantities, and then places orders with the backup supplier. Specifically, the firm first decides to reserve a capacity of $\bar{Q}^f$ units from the secondary flexible backup supplier and pays $u^f \times \bar{Q}^f$ to do so. Simultaneously, the firm places orders with the dedicated unreliable suppliers (Stage 1). If the corresponding supplier is up, the orders are fully delivered, and the firm pays the full purchasing price to the supplier. Otherwise, nothing is delivered, and therefore, the firm does not pay the purchasing price (i.e., as before, the firm pays only per item delivered). The firm can then use its flexible backup supplier subject to the reserved capacity, $\bar{Q}^f$, by paying the capacity reservation exercise prices (Stage 2). Then demands are realized and either inventory shortage cost or holding minus the sales revenue accrue.

Let $C_{U,U}(\bar{Q}^f, q^1, q^2)$, $C_{U,D}(\bar{Q}^f, q^1)$, $C_{D,U}(\bar{Q}^f, q^2)$, and $C_{D,D}(\bar{Q}^f)$ (U: Up, D: Down) denote the minimum expected cost of the firm in the second stage, if both suppliers deliver (i.e., has been up), only the first supplier delivers, only the second supplier

delivers, and when none of them delivers, respectively. Analogous to (5.3)-(5.6), these costs can be computed as follows.

$$\begin{aligned} C_{U,U}(\bar{Q}^f, q^1, q^2) &= c^1\, q^1 + c^2\, q^2 \\ &+ \min_{q_1^f, q_2^f \geq 0 \,\text{s.t.}\, q_1^f + q_2^f \leq \bar{Q}^f} \sum_{j=1}^{2} c_j^f\, q_j^f + G_1(q^1 + q_1^f) + G_2(q^2 + q_2^f) \quad (5.12) \\ C_{U,D}(\bar{Q}^f, q^1) &= c^1\, q^1 + \min_{q_1^f, q_2^f \geq 0 \,\text{s.t.}\, q_1^f + q_2^f \leq \bar{Q}^f} \sum_{j=1}^{2} c_j^f\, q_j^f + G_1(q^1 + q_1^f) + G_2(q_2^f) \quad (5.13) \\ C_{D,U}(\bar{Q}^f, q^2) &= c^2\, q^2 + \min_{q_1^f, q_2^f \geq 0 \,\text{s.t.}\, q_1^f + q_2^f \leq \bar{Q}^f} \sum_{j=1}^{2} c_j^f\, q_j^f + G_1(q_1^f) + G_2(q^2 + q_2^f) \quad (5.14) \\ C_{D,D}(\bar{Q}^f) &= \min_{q_1^f, q_2^f \geq 0 \,\text{s.t.}\, q_1^f + q_2^f \leq \bar{Q}^f} \sum_{j=1}^{2} c_j^f\, q_j^f + G_1(q_1^f) + G_2(q_2^f). \quad (5.15) \end{aligned}$$

Next, if $C_{Stage\,2}(\bar{Q}^f, q^1, q^2)$ denotes the optimal expected cost of Stage 2 as is perceived by the firm at the beginning of Stage 1, we have: $C_{Stage\,2}(\bar{Q}^f, q^1, q^2) =$

$$\begin{aligned} &\theta^1\,\theta^2\, C_{U,U}(\bar{Q}^f, q^1, q^2) + \theta^1\,(1-\theta^2\,) C_{U,D}(\bar{Q}^f, q^1) \\ &+(1-\theta^1)\,\theta^2\, C_{D,U}(\bar{Q}^f, q^2) + (1-\theta^1)(1-\theta^2)\, C_{D,D}(\bar{Q}^f). \quad (5.16) \end{aligned}$$

Then, the firm can determine $\bar{Q}^{f*}$ as well as q^{1*}, q^{2*} at the beginning of Stage 1 via:

$$\min_{q^1, q^2, \bar{Q}^f \geq 0} \quad u^f\, \bar{Q}^f + C_{Stage\,2}(\bar{Q}^f, q^1, q^2). \quad (5.17)$$

Notice that the solution to (5.15) is already given in Theorem 9, although the value of $\bar{Q}^f$ may differ. Hence, we now need to solve programs (5.12) - (5.14).

Theorem 11 (Both Suppliers Up) *For $j = 1, 2$, let $\alpha_j = [F_j^{-1}(\frac{p_j + r_j - c_j^f}{p_j + r_j + h_j}) - q^j]^+$ and $\beta_j = [F_j^{-1}([\frac{p_j + r_j - c_j^f - t}{p_j + r_j + h_j}]^+) - q^j]^+$, where t is the solution to $\sum_{j=1}^{2} \beta_j = \bar{Q}^f$. If $\bar{Q}^f \in \big[\sum_{j=1}^{2} \alpha_j\,, \infty\big)$, then $q_j^{f*} = \alpha_j$. Otherwise, $q_j^{f*} = \beta_j$.*

The above theorem states that, if the reserved capacity from the backup supplier is enough, the firm will set the order-up-to level of product j equal to $F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j})$; otherwise, to set the order-up-to level, the firm rations the available backup capacity between the two products using parameter t. Here, t can be thought of an additional fictitious capacity reservation exercise cost that rations the limited capacity. In each of these cases, and for both products, if the delivered order from the dedicated supplier is more than the order-up-to level, the firm will not use the backup capacity. Otherwise, the firm brings its inventory level to the order-up-to level by reordering the rest of its requirement from the backup supplier and paying the capacity reservation exercise cost.

Next, we solve programs (5.13) and (5.14) to analyze the case where exactly one of the dedicated suppliers fails to deliver.

Theorem 12 (One Supplier Up) *Let $m \in \{1,2\}$ denote the dedicated supplier that delivers, and $n = 3 - m$ be the disrupted supplier. Let $\alpha_m = [F_m^{-1}(\frac{p_m+r_m-c_m^f}{p_m+r_m+h_m}) - q^m]^+$, $\alpha_n = F_n^{-1}(\frac{p_n+r_n-c_n^f}{p_n+r_n+h_n})$, $\beta_m = [F_m^{-1}([\frac{p_m+r_m-c_m^f-t}{p_m+r_m+h_m}]^+) - q^m]^+$, and $\beta_n = F_n^{-1}([\frac{p_n+r_n-c_n^f-t}{p_n+r_n+h_n}]^+)$, where t is the solution to $\sum_{j=m,n}\beta_j = \bar{Q}^f$. If $\bar{Q}^f \in \big[\sum_{j=m,n}\alpha_j\,,\,\infty\big)$, then $q_j^{f*} = \alpha_j$ for $j \in \{m,n\}$. Otherwise, $q_j^{f*} = \beta_j$.*

The above result states that the firm will set the order-up-to levels as if both unreliable suppliers have delivered. However, unlike the case where both suppliers deliver, the firm can only reach the order-up-to level for product n via the flexible backup supplier. Hence, for instance, if the reserved capacity is enough, the firm will always order a positive amount from the backup supplier for product n. When the reserved backup capacity is not enough, the firm will ration the available limited backup capacity, considering the amount delivered for product m.

Since the solution to (5.15) is already given in Theorem 9, we have completely solved programs (5.12) - (5.15), and hence, $C_{Stage\,2}(\bar{Q}^f, q^1, q^2)$ is completely computed. It is then straightforward to solve (5.17) to characterize the firm's behavior in the

first stage. Furthermore, it should be clear that because of the early orders placed with the unreliable suppliers, the optimal cost in (5.17) provides an upper bound for the optimal cost in the previous section.

5.5 Benchmark Analyses: No Recourse

To generate insights into the value of the recourse option for firms (Questions 2 and 3) and provide some benchmark analyses for Section 5.4, we now consider the much simpler case where recourse is not allowed. We present the main results here and provide further details about this setting in Appendix C.

Theorem 13 *Without the recourse option, the perceived optimal ordering and contracting decisions for the firm with* $0 \leq \theta^j < 1$ $(j = 1, 2)$ *are:*

$$q_j^{f*} = F_j^{-1}([\frac{p_j + r_j - \frac{u_j^f - \theta^j c^j}{1-\theta^j}}{p_j + r_j + h_j}]^+) \qquad (j = 1, 2), \tag{5.18}$$

$$q^{j*} = F_j^{-1}(\frac{p_j + r_j - c^j}{p_j + r_j + h_j}) - F_j^{-1}([\frac{p_j + r_j - \frac{u_j^f - \theta^j c^j}{1-\theta^j}}{p_j + r_j + h_j}]^+) \qquad (j = 1, 2), \tag{5.19}$$

$$\bar{Q^f}^* = \sum_{j=1}^{2} q_j^{f*} = \sum_{j=1}^{2} F_j^{-1}([\frac{p_j + r_j - \frac{u_j^f - \theta^j c^j}{1-\theta^j}}{p_j + r_j + h_j}]^+). \tag{5.20}$$

Theorem 13 shows in Eq. (5.20) the amount of capacity that the firm reserves, where $u_j^f = u^f + c_j^f$. For each of the products, the firm will order in total (from both suppliers of that product) the same amount that it would order if it had a single, reliable, dedicated supplier with linear ordering cost c^j. However, its *perceived* optimal ordering quantity from the flexible supplier is modified to include its unreliability beliefs about dedicated suppliers. It will then procure the rest of its requirements from the unreliable dedicated supplier. Theorem 13 also proves a separability phenomenon in capacity reservation for the two products. As the reader may have expected, the flexible backup supplier in this benchmark setting has capacity reserved at the same

levels as if there are two separate backup suppliers. This is intuitive because the joint backup capacity is not decided a priori in this section, but is part of the optimization. However, as we observed in Section 5.4, this separability disappears where recourse is allowed.

5.5.1 Benchmark Setting: The Value of the Secondary Flexible Backup Supplier

We denote the true (and not the perceived) value of the flexible backup supplier for the firm by V^f and define it as:

$$V^f = C_T\left(q'^{1*}, q'^{2*}, 0, 0, 0\right) - C_T\left(q^{1*}, q^{2*}, q_1^{f*}, q_2^{f*}, \bar{Q}^{f*}\right), \tag{5.21}$$

where $C_T(\cdot)$ (as is defined in Eq. (5.96) of Appendix C) denotes the true expected cost of the firm under its perceived optimal decisions, and q'^{j*} represents the firm's perceived optimal ordering quantity to dedicated supplier j in the absence of the flexible supplier.

Using the perceived optimal ordering and contracting levels presented in Theorem 13, we now derive the true value of the flexible backup supplier for the firm in the following lemma.

Lemma 5 *The true value of the flexible backup supplier for the firm under the capacity reservation contract is:*

$$V^f = \sum_{j=1}^{2}[\pi_0^j\left(p_j E(D_j) - G_j(q_j^{f*})\right) - (u_j^f - (1-\pi_0^j)c^j)q_j^{f*}], \tag{5.22}$$

where $G_j(\cdot)$ and q_j^{f*} are defined in Eq's. (5.2) and (5.18) respectively.

Now that we have a measure for the value of the flexible backup supplier, we can answer an interesting question:

Question 6 *If a firm perceives the capacity reservation contract with a flexible backup supplier to be valuable (and hence will wish to form a contract), will such a contract be also* truly *valuable or not (and vice versa)?*

Theorem 14 *(i) The firm perceives the capacity reservation contract with the flexible backup supplier to be valuable, if and only if, its reliability belief vector* $\Theta = (\theta^1, \theta^2)$ *satisfies:*

$$\exists j \in \{1, 2\} : \ \theta^j < \frac{p_j + r_j - u_j^f}{p_j + r_j - c^j} \, . \tag{5.23}$$

(ii) The capacity reservation contract with the flexible backup supplier is <u>not</u> *truly valuable for the firm if for both* $j = 1, 2$*:*

$$max \ \{\theta^j, 1 - \pi_0^j\} \geq \frac{p_j + r_j - u_j^f}{p_j + r_j - c^j} \, . \tag{5.24}$$

From Theorem 14, we observe that if $\frac{p_j + r_j - u_j^f}{p_j + r_j - c^j} \leq \theta^j$ for both $j = 1, 2$, the firm is *lucky* that its belief is true (regardless of whether it is underestimating or overestimating): contracting with the flexible backup supplier is neither perceived to be valuable nor is truly valuable for the firm. Also, a firm that overestimates the reliabilities of both of dedicated suppliers (i.e, $\epsilon^j > 0$ for both $j = 1, 2$) perceives the flexible backup supplier to be valuable if, and only if, it is truly valuable (for the only if part, follow the proof of Theorem 2). In fact, we have the following observation:

Observation 9 Overestimating reliabilities does not have the danger of mismatching perception and reality regarding the decision of whether or not to reserve some flexible backup capacity.

While the above observation presents a nice property of overestimating, it does not mean that it is more profitable to overestimate the reliabilities. Indeed, we can rigorously prove the following theorem which verifies the intuitive notion that firms with a more accurate reliability belief (achievable through monitoring unreliable

suppliers) can benefit more from contracting with a flexible backup supplier in terms of actual cost reduction.

Proposition 11 *The true value of the flexible backup supplier based on the firm's errors in its reliability belief, denoted by* $V^f(\epsilon^1, \epsilon^2)$*, is non-increasing in the degree of disruption risk perception error. That is, for* $j \in \{1, 2\}$ *if* $\epsilon^j > 0$*, let* $0 \leq \delta^j < \epsilon^j$ *and if* $\epsilon^j < 0$*, let* $\epsilon^j < \delta^j \leq 0$*; then* $V^f(\epsilon^1, \epsilon^2) \leq V^f(\delta^1, \delta^2)$*.*

The (numerical) Study 4 of Appendix C generates more insights into the value of the flexible backup supplier. In particular, the following observation from this study is of interest:

Observation 10 Even with large errors in its reliability perception of the primary suppliers, a firm with sufficiently high profit margin can greatly benefit from reserving some flexible backup capacity (despite the fact that the quality of information is poor).

5.5.2 Benchmark Setting: The Value of Disruption Risk Information

As the previous section suggested, a firm gains additional benefit if it obtains disruption risk information and removes the uncertainty about the disruption risks of its suppliers. For instance, in the example of Boeing's supply chain discussed in Section 1, monitoring the production problems of AIT in 2006 could help Boeing to protect against the disruption.

Clearly, obtaining disruption risk information is costly (e.g. the cost of establishing a threat level advisory system, providing suppliers with incentives to collaboratively share their related private information, or placing some of the firm's employees at the supplier's site). Hence, there is a trade-off between the cost of obtaining such information and the savings due to better contracting and ordering decisions. To examine this trade-off, we denote by V^{ij} the value of obtaining perfect information

on dedicated supplier j given that this supplier is in state $i \in \{0, +\}$ and define it as:

$$V^{i^j} = C_T\left(q^{1*}, q^{2*}, q_1^{f*}, q_2^{f*}, \bar{Q}^{f*}\right) - C_T\left(q^{1\#}, q^{2\#}, q_1^{f\#}, q_2^{f\#}, \bar{Q}^{f\#} | i^j\right), \tag{5.25}$$

where superscript # on decision variables describes that they are the firm's perceived optimal decisions with respect to the new information (i.e., i^j), and $C_T\left(\cdot | i^j\right)$ is the true expected cost of the firm under such decisions (given that dedicated supplier j is in state i^j). The value of the information on the disruption risk of unreliable supplier j $(j = 1, 2)$ can then be computed by:

$$VI^j = P(\mathbb{1}_{(i^j=+)} = 1)\, V^{+^j} + P(\mathbb{1}_{(i^j=0)} = 1)\, V^{0^j} = (1 - \pi_0^j)\, V^{+^j} + \pi_0^j\, V^{0^j}, \tag{5.26}$$

where we have used $P(\mathbb{1}_{(i^j=0)} = 1) = \pi_0^j$. (Recall that $\Pi = (1 - \pi_0^1, 1 - \pi_0^2)$ is the vector of *true* reliabilities.) It is noteworthy that the value of information on dedicated supplier j (VI^j) defined in (5.26) also represents an upper bound for the amount of money that a risk-neutral firm should be willing to pay to obtain information about the disruption risk of dedicated supplier j.

To compute the total value of information, we can define the aggregate value of information for the firm as $VI = \sum_{j=1}^{2} VI^j$, representing the savings that the firm can obtain in its true expected costs by moving from a no-information situation to a full/perfect information one, as computed in the following lemma.

Lemma 6 *The values of the information on the disruption risk of the dedicated sup-*

plier j $(j = 1, 2)$ under the capacity reservation contract for the firm are:

$$V^{+^j} = (u_j^f - (1 - \pi_0^j)\, c^j) q_j^{f*} + \pi_0^j \left[G_j(q_j^{f*}) - G_j(q_j^{f*} + q^{j*}) - c^j(q_j^{f*} + q^{j*}) \right], \quad (5.27)$$

$$V^{0^j} = (1 - \pi_0^j) \left[c^j q^{j*} + G_j(q^{j*} + q_j^{f*}) - G_j(F_j^{-1}(\frac{p_j + r_j - u_j^f}{p_j + r_j + h_j})) \right] + \pi_0^j \left[G_j(q_j^{f*}) \right.$$
$$\left. - G_j(F_j^{-1}(\frac{p_j + r_j - u_j^f}{p_j + r_j + h_j})) \right] - u_j^f (F_j^{-1}(\frac{p_j + r_j - u_j^f}{p_j + r_j + h_j}) - q_j^{f*}), \quad (5.28)$$

$$VI^j = (u_j^f - (1 - \pi_0^j)\, c^j) q_j^{f*} + \pi_0^j \left[G_j(q_j^{f*}) - G_j(F_j^{-1}(\frac{p_j + r_j - u_j^f}{p_j + r_j + h_j})) \right.$$
$$\left. - u_j^f F_j^{-1}(\frac{p_j + r_j - u_j^f}{p_j + r_j + h_j}) \right], \quad (5.29)$$

where $G_j(\cdot)$, q_j^{f*}, and q^{j*} are defined in Eq's. (5.2), (5.18), and (5.19) respectively.

Now that we have a measure for the value of disruption risk information, we can answer two interesting questions:

Question 7 *When is obtaining disruption risk information a better risk mitigation mechanism than contracting with a flexible backup supplier?*

Question 8 *For which firms should obtaining disruption risk information be more appealing?*

We first provide insight into Question 7.

Proposition 12 *There exists thresholds $\hat{\pi}_0^j$ on the true unreliability of the primary suppliers such that $VI \geq V^f$ whenever $\pi_0^j \leq \hat{\pi}_0^j$ $(j \in \{1, 2\})$. That is, when suppliers are (truly) reliable enough, obtaining information is more valuable than contracting with a flexible backup supplier.*

When primary suppliers are reliable enough, investing in an expensive backup supplier is not advantageous. However, obtaining information is still (relatively) advantageous because it helps the firm to make better ordering decisions.

The following proposition (and later Observation 11) provide answers to Question 8. First, as is intuitively expected, firms that currently do not have an accurate vector of reliability belief can achieve larger savings in their true costs through obtaining risk information. Hence, monitoring suppliers should be more attractive for such firms.

Proposition 13 *The value of information based on the firm's errors in its reliability belief, denoted by* $VI(\epsilon^1, \epsilon^2)$*, is non-decreasing in the degree of disruption risk perception error. That is, for* $j \in \{1, 2\}$ *if* $\epsilon^j > 0$*, let* $0 \leq \delta^j < \epsilon^j$ *and if* $\epsilon^j < 0$*, let* $\epsilon^j < \delta^j \leq 0$*; then* $VI(\epsilon^1, \epsilon^2) \geq VI(\delta^1, \delta^2)$*. Furthermore, if* $|\delta^j| \geq |\epsilon^j|$*, then* $VI(\epsilon^1, \epsilon^2) \leq VI(\delta^1, \delta^2)$*.*

The (numerical) Study 5 of Appendix C provides the following observations about disruption risk information.

Observation 11 Disruption risk information is more attractive to firms with lower profit margins.

Observation 12 The sensitivity of the value of information to belief errors is much higher for firms that tend to overestimate reliabilities than those who underestimate.

5.6 The Value of Recourse, Flexibility, and Information

We now use our framework with recourse to gain insights into the value of three disruption mitigation mechanisms: obtaining (i) a recourse option, (ii) flexibility in the backup supply system, and (3) disruption risk information on primary suppliers.

5.6.1 The Value of Recourse

To reveal the benefit of recourse and provide insights to Question 3, we now compare the optimal cost of the model with the recourse to the benchmark setting of

Section 5.5 (where the firm cannot observe the state of the unreliable suppliers before ordering).

Study 1 (Recourse) As in Studies 4 and 5 (described in details in Appendix C), consider a firm which is facing Normally distributed demands of $N(5000, 1200^2)$ and $N(3000, 800^2)$, respectively, for products 1 and 2. Table 5.1 describes the percentage benefit of recourse as well as the percentage reduction in optimal investment in the flexible capacity based on the parameter settings of Table 5.4 (see Appendix B). To focus on the effect of recourse, we assume (1) the firm has no error in its risk belief, and (2) c_j^f $(j = 1, 2)$ is negligible compared to u^f. By comparing the settings with recourse and without it in Table 5.1, we gain the following insights into Questions 2 and 3.

Observation 13 Recourse is a strong mitigation technique for firms with a cost reduction that is always positive and averages 37.7% in our study.

Since this reduction in cost is due to making more effective procurement decisions as a consequence of observing disruption states, the above analysis gives quantitative evidence of the value of postponing ordering decisions (to the extent possible) until after monitoring the state of unreliable suppliers. Another interesting observation is the following.

Observation 14 Investment in the flexible backup capacity may be greater or smaller with recourse.

One might think that a firm with recourse would always invest less in the flexible capacity, since later it can benefit from its disruption observation to flexibly utilize the reserved pooled capacity (a *risk pooling effect*). However, this is not true in Settings 3, 4, and 7. For instance, in Setting 4, the firm without recourse, does not invest in the flexible capacity (according to Theorem 14). With a recourse option, however, it knows that if in the second stage it observes that (at least) one of the suppliers

Table 5.1: The Value of Recourse and the Difference in Investment in the Flexible Backup Capacity with and without Recourse.

Setting No.	**Recourse**		**No Recourse**		Reduction in $\bar{Q}^{f*}$ (%)	Value of Recourse (%)
	Q^{f*}	Opt. Cost	Q^{f*}	Opt. Cost		
1	3,314	-6,766.7	6,845	-3,156.0	51.59	114.41
2	3,287	-6,673.4	5,415	-3,563.2	39.30	87.29
3	3,214	-7,196.0	2,571	-4,771.7	-25.01	50.81
4	2,459	-7,495.7	0	-7,050.8	$-\infty$	6.31
5	5,206	-93,662.7	8,308	-90,893.4	37.34	3.05
6	5,095	-26,510.3	8,495	-25,588.9	40.02	3.60
7	3,233	-26,484.6	2,436	-23,940.9	-32.72	10.62
8	3,520	-24,025.0	6432	-19,086.8	45.27	25.87
Average						**37.744**

is down, it has much to gain by channeling the reserved capacity to the appropriate product(s). Hence, it prefers to invest in the secondary supplier to reduce the risk.

5.6.2 The Value of Flexibility

We now use our analytical framework to provide insights into Question 4.

Study 2 (Flexibility) To capture the value of implementing flexibility in the backup system, we compare two scenarios: (1) one with two dedicated (i.e., inflexible) backup suppliers, where the dedicated backup supplier of product j has a capacity reservation cost of u_j^d, and (2) one with a single flexible backup supplier. Notice that, for computational purposes, the first scenario is a special case of our modeling framework, and to analyze it, one can simply use the results provided in Section 5.4 twice (i.e., separately for each product), each time setting the demand for one of the products to 0. To investigate the value of flexibility, we assume $u_j^d = (1+\Delta)c^j$, where Δ represents a "backup premium." Then, to fairly price the capacity of the flexible supplier, we consider u^f as a weighted average (based on quantities demanded for each) of u_1^d and u_2^d, and set $u^f = [\sum_{j=1,2} E(D_j)\,\bar{\theta}^j\,u_j^d]/[\sum_{j=1,2} E(D_j)\,\bar{\theta}^j]$. The other parameter settings and assumptions are the same as those used in Study 1. The results presented in Table 5.2 lead to the following two observations.

Observation 15 The (mix) flexibility of a backup supplier is highly beneficial to a

Table 5.2: The Value of Flexibility and the Reduction in Investment in the Backup Capacity due to Flexibility.

Setting No.	**Value of Flexibility (%)**					**Reduction in Total Capacity Reserved (%)**				
	Δ					Δ				
	0%	5%	10%	15%	AVG. (%)	0%	5%	10%	15%	AVG. (%)
1	0.5	16.6	28.0	44.0	22.3	31.6	36.6	16.6	36.0	30.2
2	8.3	16.5	31.6	36.2	23.2	24.9	45.2	16.5	39.6	31.5
3	4.6	13.1	24.9	37.5	20.0	22.3	46.7	13.1	55.8	34.5
4	6.8	16.0	22.3	17.6	15.7	1.6	49.7	16.0	5.1	18.1
5	0.5	1.0	1.6	2.1	1.3	9.3	13.5	1.0	36.4	15.0
6	0.0	3.6	9.3	9.9	5.7	104.9	0.0	3.6	40.3	37.2
7	101.7	103.2	104.9	107.4	104.3	104.0	-54.1	103.2	-48.6	26.1
8	96.7	97.4	104.0	105.8	101.0	40.8	-100.6	97.4	-55.7	-4.5
AVG. (%)	27.4	33.4	40.8	45.1	**36.78**	42.4	4.6	33.4	13.6	**23.52**

firm that is procuring from unreliable primary suppliers, with an average cost reduction of 36.7% in our study. Moreover, the value of implementing flexibility in the backup system increases as the backup premium increases.

Observation 16 The capacity reserved with a single flexible backup supplier is not always less than the total capacity reserved with two dedicated backup suppliers. However, the flexibility results in an average reduction of 23.5% (in our study) in the total backup capacity purchased.

5.6.3 The Value of Disruption Risk Information

Finally, we use our analytical framework to gain insights into Question 5 of the introduction to Section 4.

Study 3 (Information) Consider the parameter settings of Study 1 (Table 5.4 of Appendix B), but assume the firm's disruption risk belief (θ^1, θ^2) is subject to errors as presented in Table 5.3. Table 5.3 presents the value of (perfect) information by comparing the cost of the firm when its decision is based only on its belief (Imperfect Information) with a scenario where it obtains risk information and decides based on the true risk of its suppliers (Perfect Information). From Table 5.3 we see that obtaining information may decrease or increase the firm's investment level in flexible backup

Table 5.3: The Value of Disruption Risk Information Under Recourse.

Setting	Belief Error		**Perfect Info.**		**Imperfect Info.**		Reduction in $\bar{Q}^{f*}$(%)	Value of Info. (%)
	ϵ^1	ϵ^2	$\bar{Q}^{f*}$	Opt. Cost	$\bar{Q}^{f*}$	Opt. Cost		
1	0.1	0.1	3,886	-7,028.4	3,314	-6,956.7	-17.26	1.03
2	-0.1	-0.1	0	-7,861.4	3,287	-7,334.1	∞	7.19
3	0.1	-0.1	3,965	-6,136.1	3,214	-6,003.4	-23.40	2.21
4	0.2	0.15	3,924	-6,581.8	2,459	-5,701.4	-59.58	15.44
5	0.15	0.2	7,463	-91,331.9	5,206	-90,117.2	-43.37	1.35
6	-0.1	0.2	3,755	-26,660.9	5,095	-26,548.1	26.29	0.42
7	0.2	0.2	4,664	-26,504.6	3,233	-25,654.2	-44.29	3.31
8	0.15	0.2	4,486	-24,457.9	3,520	-23,977.1	-27.46	2.01
Average								**4.121**

capacity, depending on whether the firm has been overestimating or underestimating the risks. Furthermore, we gain the following insight into Question 8.

Observation 17 Obtaining perfect disruption risk information with a recourse option results in an average cost reduction of 4.12%, and it is not a very strong risk mitigation technique compared to obtaining a recourse option and/or implementing flexibility in the backup system.

In other words, once the firm obtains a recourse option to benefit from the flexible backup capacity, the additional benefit of reducing risk belief errors is modest. Additional insights regarding the role of profit margin on the value of disruption information are found in Appendix C.

5.7 Summary of Findings and Conclusion

We developed a rigorous quantitative methodology to capture the value of two key supply risk mitigation mechanisms: (1) contracting with a secondary flexible backup supplier, and (2) obtaining disruption risk information through monitoring primary suppliers. We derived analytical measures for the true value of a flexible backup supplier as well as the value of obtaining disruption risk information. These measures determine upper bounds for the amount of money that a risk-neutral firm should

be willing to invest to implement either of these strategies in order to increase the reliability and responsiveness of its supply chain.

In both settings with and without a recourse option, we analytically characterized the firm's behavior by explicitly identifying the jointly optimal size of the backup capacity reservation contract and the inventory ordering policy for both products. This characterization was based upon the firm's perception of the primary suppliers' disruption risks.

We observed that investing in a secondary flexible backup capacity can be harmful if the current information about the risk of primary suppliers is not perfect. We showed that monitoring unreliable suppliers to make better risk estimates enhances the benefit of purchasing flexible backup capacity. We also identified conditions under which a firm is lucky in the sense that regardless of whether it is overestimating or underestimating the reliabilities, it perceives investing in a flexible backup capacity to be valuable only if it is truly valuable. For instance, we found that overestimating supplier reliabilities does not have the danger of mismatching perception and reality regarding a decision to reserve some flexible backup capacity. Moreover, we showed that contracting with a flexible backup supplier is more beneficial for firms with low perception errors about the reliability of their suppliers than those with high errors. By contrast, disruption risk information is more valuable for firms with higher perception errors. Additionally, we observed that disruption risk information is more attractive to firms with low profit margins than those with high ones. We also showed that when suppliers are (truly) reliable enough, obtaining information is a better risk mitigation technique than contracting with a flexible supplier. We also found that the value of disruption risk information is much more sensitive to the misperception errors for firms who tend to overestimate (rather than underestimate) the reliabilities.

Next, comparing the scenarios with and without the recourse option, our study found that having the recourse option can be regarded as an effective risk mitigation technique for firms with an average cost reduction of 37%. This observation sheds

more light on the benefit of monitoring suppliers and provides further evidence that firms with unreliable suppliers should try to postpone (to the extent possible) their ordering decisions until after monitoring the disruption state of their suppliers. We also observed that the amount of investment in the flexible backup capacity may or may not be reduced when a firm obtains a recourse option. Further, we showed that when the perceived reliability of the suppliers is larger than a critical fraction, having a recourse option reduces the optimal investment in the flexible backup capacity.

We investigated the value of implementing flexibility in the backup system: contracting with a single flexible backup supplier rather than two inflexible ones. Our study showed an average cost reduction of 36%, so flexibility can indeed be highly beneficial; further, it becomes more beneficial as the backup premium increases.

We evaluated the benefit of obtaining risk information under the recourse option, but our results suggest that it is not a strong mitigation mechanism. Without recourse, it is potent. We also extended our two-stage analyses to a similar setting with offshore unreliable suppliers, where the availability of unreliable suppliers can only be identified by observing the delivered quantities.

We leave it to future research to investigate the effects of dynamic changes in the reliability of the suppliers on the results provided in this study. Future research may also investigate the multi-period trade-offs in carrying inventory over time and dynamically monitoring suppliers to hedge against such dynamically changing risks. Another possible direction for future research is to investigate the effect of correlation in disruption risk across different suppliers and/or correlation in demand across different products.

5.8 Appendix A: Proofs

Proof of Proposition 8: Note that the KKT conditions are sufficient and necessary for characterizing the optimal solution (i.e., the optimal ordering policy). Let μ, λ^f, and λ denote the Lagrangian multipliers for constraints $q^f \leq \bar{Q^f}$, $q^f \geq 0$, and $q \geq 0$, respectively. First consider the case where the unreliable supplier is observed to be up. Using the Leibniz rule to get the derivative of the objective function, $c^f q^f + cq + G(q + q^f)$, the KKT conditions can be written as:

$$
\begin{aligned}
& q^f \leq \bar{Q^f} \\
& c + (h + p + r)F(q + q^f) - p - r = \lambda \\
& c^f + (h + p + r)F(q + q^f) - p - r = \lambda^f - \mu \\
& \mu(q^f - \bar{Q^f}) = 0 \\
& \lambda q = 0 \\
& \lambda^f q^f = 0 \\
& q, q^f, \mu, \lambda, \lambda^f \geq 0,
\end{aligned}
$$

or equivalently

$$
\begin{aligned}
& q^f \leq \bar{Q^f} \\
& c - \lambda = \mu - \lambda^f + c^f \\
& q + q^f = F^{-1}(\frac{p + r - (c - \lambda)}{p + r + h}) \\
& \mu(q^f - \bar{Q^f}) = 0 \\
& \lambda q = 0 \\
& \lambda^f q^f = 0 \\
& q, q^f, \mu, \lambda, \lambda^f \geq 0,
\end{aligned}
$$

Similarly, when the unreliable supplier is down, the KKT conditions are:

$$
\begin{aligned}
& q^f \leq \bar{Q^f} \\
& q^f = F^{-1}(\frac{p+r-(c^f+\mu-\lambda^f)}{p+r+h}) \\
& \mu(q^f - \bar{Q^f}) = 0 \\
& \lambda^f q^f = 0 \\
& q^f, \mu, \lambda^f \geq 0.
\end{aligned}
$$

To prove Part (1), set $q^{f*} = \mu = \lambda = 0, \lambda^f = c^f - c$, and $q^* = F^{-1}(\frac{p+r-c}{p+r+h})$, and observe that the KKT conditions are satisfied. To prove Part (2), if $\bar{Q^f} \geq F^{-1}(\frac{p+r-c^f}{p+r+h})$, then set $q^{f*} = F^{-1}(\frac{p+r-c^f}{p+r+h})$, $\mu = \lambda^f = 0$, and observe that KKT conditions are satisfied. Otherwise, choose μ such that $\bar{Q^f} = F^{-1}(\frac{p+r-(c^f+\mu)}{p+r+h})$. Then observe that setting $q^{f*} = \bar{Q^f}$ and $\lambda^f = 0$ satisfies the KKT conditions. For Part (3) (i), set $\mu = \lambda^f = 0$, $\lambda = c - c^f$, $q^* = 0$, and $q^{f*} = F^{-1}(\frac{p+r-c^f}{p+r+h})$. For Part (3) (ii), choose μ such that $\bar{Q^f} = F^{-1}(\frac{p+r-(c^f+\mu)}{p+r+h})$ and set $\lambda^f = 0$, $\lambda = c - c^f + \mu$, $q^{f*} = \bar{Q^f}$, and $q^* = 0$. For Part (3) (iii), set $\mu = c - c^f$, $\lambda^f = \lambda = 0$, $q^{f*} = \bar{Q^f}$, $q^* = F^{-1}(\frac{p+r-c}{p+r+h}) - \bar{Q^f}$. Similarly, to prove Part (4) (i), set $\mu = \lambda^f = 0$ and $q^{f*} = F^{-1}(\frac{p+r-c^f}{p+r+h})$, and to prove Part (4) (ii), choose μ such that $\bar{Q^f} = F^{-1}(\frac{p+r-(c^f+\mu)}{p+r+h})$ and set $\lambda^f = 0$, and $q^{f*} = \bar{Q^f}$. □

Proof of Proposition 9: Recall that the optimal capacity reservation level, similar to 5.8, is $\bar{Q^{f*}} = \arg\min_{\bar{Q^f} \geq 0} u^f \bar{Q^f} + \theta C_U(\bar{Q^f}) + (1-\theta) C_D(\bar{Q^f})$, where $C_U(\cdot)$ ($C_D(\cdot)$) denotes the cost of Stage 2 if the unreliable supplier is observed to be up (down). For the ease of notation, let $l = F^{-1}(\frac{p+r-c^f}{p+r+h})$ and $l' = F^{-1}(\frac{p+r-c}{p+r+h})$. When $c^f > c$, from Proposition 8 Part (1), $C_U(\cdot) = a$ for some constant a. Moreover, from Proposition 8 Part (2), $C_D(\bar{Q^f}) = c^f(l \wedge \bar{Q^f}) + G(l \wedge \bar{Q^f})$, where $x \wedge y = \min\{x, y\}$. Thus, the optimization problem when $c^f > c$ is $\min_{\bar{Q^f} \geq 0} u^f \bar{Q^f} + (1-\theta)[c^f(l \wedge \bar{Q^f}) + G(l \wedge \bar{Q^f})]$. Note that since the optimizer of $u^f \bar{Q^f} + (1-\theta)[c^f \bar{Q^f} + G(\bar{Q^f})]$ on $\bar{Q^f} \geq 0$ is $l = F^{-1}([\frac{p+r-(u^f/(1-\theta)+c^f)}{p+r+h}]^+)$, which is always less than, the proof of Part (1) is

complete. When $c^f \leq c$, we need to consider three cases: (a) $\bar{Q^f} \in [l, \infty)$, (b) $\bar{Q^f} \in [l', l]$, and (c) $\bar{Q^f} \in [0, l']$. In case (a), from Proposition 8 Parts (3) (i) and (4) (i), we have $C_U(\bar{Q^f}) = C_D(\bar{Q^f}) = c^f l + G(l)$. Thus, in this case, the optimization problem is $\min_{\bar{Q^f} \geq l} u^f \bar{Q^f} + c^f l + G(l)$, which has the solution $\bar{Q^{f*}} = l$. In case (b), from Proposition 8 Parts (3) (ii) and (4) (ii), $C_U(\bar{Q^f}) = C_D(\bar{Q^f}) = c^f \bar{Q^f} + G(\bar{Q^f})$. Hence, the optimization problem is $\min_{l' \leq \bar{Q^f} \leq l} u^f \bar{Q^f} + c^f \bar{Q^f} + G(\bar{Q^f})$. Note that the unconstrained version of this problem has the optimizer $F^{-1}([\frac{p+r-(u^f+c^f)}{p+r+h}]^+) \leq l'$, where the inequality holds since by our modeling assumption $u^f + c^f > c$. Hence, the optimizer in case (b) is $\bar{Q^{f*}} = l'$, since the objective function is convex. In case (c), from Proposition 8 Parts (3) (iii) and (4) (ii), $C_U(\bar{Q^f}) = c^f \bar{Q^f} + G(l') + c(l' - \bar{Q^f})$ and $C_D(\bar{Q^f}) = c^f \bar{Q^f} + G(\bar{Q^f})$. Thus, the optimization problem in this case is equivalent to $\min_{0 \leq \bar{Q^f} \leq l'} (u^f + c^f - \theta c)\bar{Q^f} + (1-\theta) G(\bar{Q^f})$, which has the optimizer $F^{-1}([\frac{p+r-\frac{u^f+c^f-\theta c}{1-\theta}}{p+r+h}]^+) \leq l'$, where the inequality holds since by our modeling assumption $u^f + c^f > c$. Next observe that the optimizer in case (a) is a feasible point in case (b), and also, the optimizer of case (b), is a feasible point in case (c). Hence, the optimizer of case (c) gives the global optimal solution to the problem, and the proof is complete. □

Proof of Theorem 7: Consider program (5.3). Notice that it is a convex program with linear constraints. Hence, the KKT conditions are sufficient and necessary. Using Leibniz rule these conditions are as follows:

$$q_1^f + q_2^f \leq \bar{Q}^f$$

$$c^j + (h_j + p_j + r_j)F_j(q^j + q_j^f) - p_j - r_j = \lambda^j \qquad (j = 1, 2) \qquad (5.30)$$

$$c_j^f + (h_j + p_j + r_j)F_j(q^j + q_j^f) - p_j - r_j = \lambda_j^f - \mu \qquad (j = 1, 2) \qquad (5.31)$$

$$\mu(q_1^f + q_2^f - \bar{Q}^f) = 0$$

$$\lambda^j q^j = 0 \qquad (j = 1, 2)$$

$$\lambda_j^f q_j^f = 0 \qquad (j = 1, 2)$$

$$q^j, q_j^f, \mu, \lambda^j, \lambda_j^f \geq 0. \qquad (j = 1, 2)$$

Conditions (5.30) and (5.31) result that:

$$c^j - \lambda^j = \mu - \lambda_j^f + c_j^f \qquad (j = 1, 2)$$

$$q^j + q_j^f = F_j^{-1}(\frac{p_j + r_j - (c^j - \lambda^j)}{p_j + r_j + h_j}). \qquad (j = 1, 2)$$

Hence, the KKT conditions can be written as follows:

$$q_1^f + q_2^f \leq \bar{Q}^f \qquad (5.32)$$

$$c^j - \lambda^j = \mu - \lambda_j^f + c_j^f \qquad (j = 1, 2) \qquad (5.33)$$

$$q^j + q_j^f = F_j^{-1}(\frac{p_j + r_j - (c^j - \lambda^j)}{p_j + r_j + h_j}) \qquad (j = 1, 2) \qquad (5.34)$$

$$\mu(q_1^f + q_2^f - \bar{Q}^f) = 0 \qquad (5.35)$$

$$\lambda^j q^j = 0 \qquad (j = 1, 2) \qquad (5.36)$$

$$\lambda_j^f q_j^f = 0 \qquad (j = 1, 2) \qquad (5.37)$$

$$q^j, q_j^f, \mu, \lambda^j, \lambda_j^f \geq 0. \qquad (j = 1, 2) \qquad (5.38)$$

Now, it is sufficient to show that the optimal solution in each part satisfies the above conditions.

(i) If $\bar{Q}^f \in \left[\sum_{j=k,l} F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j}) \, , +\infty \right)$ set $\mu = \lambda_j^f = 0$ and $\lambda^j = c^j - c_j^f$ $(j = 1, 2)$. Then observe that $q_j^{f*} = F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j})$ and $q^{j*} = 0$ $(j = 1, 2)$ satisfy the KKT conditions.

(ii) If $\bar{Q}^f \in \left[\sum_{j=k,l} F_j^{-1}(\frac{p_j+r_j-c_j^f-(c^l-c_l^f)}{p_j+r_j+h_j}) \, , \sum_{j=k,l} F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j}) \right)$, set $\mu = t, \lambda^k = c^k - c_k^f - \mu, \lambda^l = c^l - c_l^f - \mu$ and $\lambda_k^f = \lambda_l^f = 0$. Then observe that $q_j^{f*} = F_j^{-1}(\frac{p_j+r_j-c_j^f-t}{p_j+r_j+h_j})$ and $q^{j*} = 0$ (for both $j = k, l$) satisfy the KKT conditions, where $t \in (0, c^l - c_l^f]$ is a solution to: $\sum_{j=k,l} F_j^{-1}(\frac{p_j+r_j-c_j^f-t}{p_j+r_j+h_j}) = \bar{Q}^f$.

(iii) If $\bar{Q}^f \in \left[F_k^{-1}(\frac{p_k+r_k-c_k^f-(c^l-c_l^f)}{p_k+r_k+h_k}) \, , \sum_{j=k,l} F_j^{-1}(\frac{p_j+r_j-c_j^f-(c^l-c_l^f)}{p_j+r_j+h_j}) \right)$, set $\mu = c^l - c_l^f$, $\lambda_l^f = \lambda_k^f = \lambda^l = 0$ and $\lambda^k = (c^k - c_k^f) - (c^l - c_l^f)$. Then observe that $q_k^{f*} = F_k^{-1}(\frac{p_k+r_k-c_k^f-(c^l-c_l^f)}{p_k+r_k+h_k})$, $q_l^{f*} = \bar{Q}^f - q_k^{f*}$, $q^{k*} = 0$, and $q^{l*} = F_l^{-1}(\frac{p_l+r_l-c^l}{p_l+r_l+h_l}) - q_l^{f*}$ satisfy the KKT conditions.

(iv) If $\bar{Q}^f \in \left[F_k^{-1}(\frac{p_k+r_k-c^k}{p_k+r_k+h_k}) \, , F_k^{-1}(\frac{p_k+r_k-c_k^f-(c^l-c_l^f)}{p_k+r_k+h_k}) \right)$, suppose $t \in (c^l - c_l^f, c^k - c_k^f]$ is a solution to $\bar{Q}^f = F_k^{-1}(\frac{p_k+r_k-c_k^f-t}{p_k+r_k+h_k})$. Then set $\mu = t$, $\lambda^k = c^k - c_k^f - \mu$, $\lambda^l = \lambda_k^f = 0$ and $\lambda_l^f = \mu - (c^l - c_l^f)$. Then observe that $q_k^{f*} = \bar{Q}^f$, $q_l^{f*} = 0$, $q^{k*} = 0$, and $q^{l*} = F_l^{-1}(\frac{p_l+r_l-c^l}{p_l+r_l+h_l})$ satisfy the KKT conditions.

(v) If $\bar{Q}^f \in \left[0 \, , F_k^{-1}(\frac{p_k+r_k-c^k}{p_k+r_k+h_k}) \right)$, set $\mu = c^k - c_k^f$, $\lambda^l = \lambda^k = \lambda_k^f = 0$, and $\lambda_l^f = (c^k - c_k^f) - (c^l - c_l^f)$. Then observe that $q_k^{f*} = \bar{Q}^f$, $q_l^{f*} = 0$, $q^{k*} = F_k^{-1}(\frac{p_k+r_k-c^k}{p_k+r_k+h_k}) - \bar{Q}^f$, and $q^{l*} = F_l^{-1}(\frac{p_l+r_l-c^l}{p_l+r_l+h_l})$ satisfy the KKT conditions. □

Proof of Theorem 8: Consider program (5.4) or (5.5) depending on whether $m = 1$ or 2 (respectively). Notice that both programs are convex with linear constraints. Hence, the KKT conditions are sufficient and necessary. Using Leibniz rule and similar to the derivation of the KKT conditions in the proof of Theorem (7) these

conditions are as follows (see the proof of Theorem 1 for more details on the derivation of the KKT conditions).

$$
\begin{aligned}
& q_m^f + q_n^f \leq \bar{Q}^f \\
& c^m + (h_m + p_m + r_m) F_m(q^m + q_m^f) - p_m - r_m = \lambda^m && (5.39) \\
& c_m^f + (h_m + p_m + r_m) F_m(q^m + q_m^f) - p_m - r_m = \lambda_m^f - \mu && (5.40) \\
& c_n^f + (h_n + p_n + r_n) F_n(q_n^f) - p_n - r_n = \lambda_n^f - \mu && (5.41) \\
& \mu(q_m^f + q_n^f - \bar{Q}^f) = 0 \\
& \lambda^m q^m = 0 \\
& \lambda_m^f q_m^f = 0 \\
& \lambda_n^f q_n^f = 0 \\
& q^m, q_m^f, q_n^f, \mu, \lambda^m, \lambda_m^f, \lambda_n^f \geq 0.
\end{aligned}
$$

Conditions (5.39)-(5.41) result in:

$$
\begin{aligned}
& c^m - \lambda^m = \mu - \lambda_m^f + c_m^f \\
& q^m + q_m^f = F_m^{-1}\left(\frac{p_m + r_m - (\mu - \lambda_m^f + c_m^f)}{p_m + r_m + h_m}\right) \\
& q_n^f = F_n^{-1}\left(\frac{p_n + r_n - (\mu - \lambda_n^f + c_n^f)}{p_n + r_n + h_n}\right).
\end{aligned}
$$

Hence, the KKT conditions when only supplier $m \in \{1, 2\}$ turns out to be up can

be written as:

$$q_1^f + q_2^f \leq \bar{Q^f} \tag{5.42}$$

$$c^m - \lambda^m = \mu - \lambda_m^f + c_m^f \tag{5.43}$$

$$q^m + q_m^f = F_m^{-1}(\frac{p_m + r_m - (\mu - \lambda_m^f + c_m^f)}{p_m + r_m + h_m}) \tag{5.44}$$

$$q_n^f = F_n^{-1}(\frac{p_n + r_n - (\mu - \lambda_n^f + c_n^f)}{p_n + r_n + h_n}) \tag{5.45}$$

$$\lambda^m q^m = 0 \tag{5.46}$$

$$\lambda_m^f q_m^f = 0 \tag{5.47}$$

$$\lambda_n^f q_n^f = 0 \tag{5.48}$$

$$q^m, q_m^f, q_n^f, \mu, \lambda^m, \lambda_m^f, \lambda_n^f \geq 0. \tag{5.49}$$

Now, it is sufficient to show that the optimal solution in each part satisfies the above conditions.

(i) If $\bar{Q^f} \in \left[\sum_{j=n,m} F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j}), +\infty \right)$, set $\lambda^m = c^m$, $\mu = \lambda_m^f = \lambda_n^f = 0$ and observe that $q_j^{f*} = F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j})$ (for both $j = n, m$) and $q^{m*} = 0$ satisfy the KKT conditions. □

(ii) If $\bar{Q^f} \in \left[\sum_{j=n,m} F_j^{-1}([\frac{p_j+r_j-c_j^f-(c^m-c_m^f)}{p_j+r_j+h_j}])^+), \sum_{j=n,m} F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j}) \right)$, set $\mu = t$, $\lambda^m = c^m - c_m^f - t$ and $\lambda_m^f = \lambda_n^f = 0$. Then observe that $q_j^{f*} = F_j^{-1}([\frac{p_j+r_j-c_j^f-t}{p_j+r_j+h_j}]^+)$ (for both $j = n, m$) and $q^{m*} = 0$ satisfy the KKT conditions, where $t \in (0, c^m - c_m^f]$ is a solution to: $\sum_{j=n,m} F_j^{-1}([\frac{p_j+r_j-c_j^f-t}{p_j+r_j+h_j}]^+) = \bar{Q^f}$.

(iii) If $\bar{Q^f} \in \left[F_n^{-1}([\frac{p_n+r_n-c_n^f-(c^m-c_m^f)}{p_n+r_n+h_n}]^+), \sum_{j=n,m} F_j^{-1}([\frac{p_j+r_j-c_j^f-(c^m-c_m^f)}{p_j+r_j+h_j}]^+) \right)$, set $\mu = c^m - c_m^f$ and $\lambda_m^f = \lambda_n^f = \lambda^m = 0$. Then observe that $q_n^{f*} = F_n^{-1}([\frac{p_n+r_n-c_n^f-(c^m-c_m^f)}{p_n+r_n+h_n}]^+)$, $q_m^{f*} = \bar{Q^f} - q_n^{f*}$, and $q^{m*} = F_m^{-1}(\frac{p_m+r_m-c^m}{p_m+r_m+h_m}) - q_m^{f*}$ satisfy the KKT conditions.

(iv) If $\bar{Q}^f \in \left[0\,,\, F_n^{-1}([\frac{p_n+r_n-c_n^f-(c^m-c_m^f)}{p_n+r_n+h_n}]^+)\right)$ let $t \in (c^m - c_m^f, \infty)$ be a solution to $F_n^{-1}([\frac{p_n+r_n-c_n^f-t}{p_n+r_n+h_n}]^+) = \bar{Q}^f$. Set $\mu = t$, $\lambda_m^f = \mu - (c^m - c_m^f)$ and $\lambda_n^f = \lambda^m = 0$. Then observe that $q_n^{f*} = \bar{Q}^f$, $q_m^{f*} = 0$, and $q^{m*} = F_m^{-1}(\frac{p_m+r_m-c^m}{p_m+r_m+h_m})$ satisfy the KKT conditions.

□

Proof of Theorem 9: Consider program (5.6). Notice that this program is convex with linear constraints. Hence, the KKT conditions are sufficient and necessary. Using Leibniz rule and similar to the derivation of the KKT conditions in the proof of Theorems (7) and Theorem (8), these conditions are as follows:

$$
\begin{aligned}
& q_1^f + q_2^f \le \bar{Q}^f \\
& c_j^f + (h_j + p_j + r_j)F_j(q_j^f) - p_j - r_j = \lambda_j^f - \mu \qquad (j = 1, 2) \qquad (5.50) \\
& \mu(q_1^f + q_2^f - \bar{Q}^f) = 0 \\
& \lambda_j^f q_j^f = 0 \qquad (j = 1, 2) \\
& q_j^f, \mu, \lambda_j^f \ge 0. \qquad (j = 1, 2)
\end{aligned}
$$

By rewriting condition (5.50), KKT conditions are:

$$
\begin{aligned}
& q_1^f + q_2^f \le \bar{Q}^f \qquad (5.51) \\
& q_j^f = F_j^{-1}(\frac{p_j + r_j - (\mu - \lambda_j^f + c_j^f)}{p_j + r_j + h_j}) \qquad (j = 1, 2) \qquad (5.52) \\
& \mu(q_1^f + q_2^f - \bar{Q}^f) = 0 \qquad (5.53) \\
& \lambda_j^f q_j^f = 0 \qquad (j = 1, 2) \qquad (5.54) \\
& q_j^f, \mu, \lambda_j^f \ge 0. \qquad (j = 1, 2) \qquad (5.55)
\end{aligned}
$$

Now, it is sufficient to show that the optimal solution of each part satisfies the above KKT conditions.

(i) If $\bar{Q}^f \in \left[\sum_{j=1,2} F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j}) \, , +\infty \right)$, set $\mu = \lambda_1^f = \lambda_2^f = 0$. Then observe that $q_j^{f*} = F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j})$ (for both $j = 1, 2$) satisfies the KKT conditions.

(ii) If $\bar{Q}^f \in \left[0 \, , \sum_{j=1,2} F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j}) \right)$ set $\mu = t$ and $\lambda_1^f = \lambda_2^f = 0$. Then observe that $q_j^{f*} = F_j^{-1}([\frac{p_j+r_j-c_j^f-t}{p_j+r_j+h_j}]^+)$ (for both $j = n, m$) satisfies the KKT conditions, where $t \in \left(0 \, , \max_j\{p_j + r_j - c_j^f\} \right]$ is a solution to: $\sum_{j=1,2} F_j^{-1}([\frac{p_j+r_j-c_j^f-t}{p_j+r_j+h_j}]^+) = \bar{Q}^f$. $\square$

Proof of Proposition 10: Let $I_7 = \left[\sum_{j=1,2} F_j^{-1}(\frac{p_j+r_j-c_j^f}{p_j+r_j+h_j}), \infty \right)$ and notice that $\bigcup_{k=1}^7 I_k = [0, \infty)$. Thus, defining

$$\bar{Q}_k^f = \underset{\bar{Q}^f \in I_k}{\operatorname{argmin}} \quad u^f \, \bar{Q}^f + C_{Stage\,2}(\bar{Q}^f),$$

we have:

$$\bar{Q}^{f*} = \underset{\bar{Q}^f \geq 0}{\operatorname{argmin}} \quad u^f \, \bar{Q}^f + C_{Stage\,2}(\bar{Q}^f) = \underset{\bar{Q}^f \in \{\bar{Q}_k^f, k=1,...,7\}}{\operatorname{argmin}} \quad u^f \, \bar{Q}^f + C_{Stage\,2}(\bar{Q}^f) \, . \quad (5.56)$$

It remains to identify $\bar{Q}_k^f$ for $k = 1, 2, ..., 7$ using Theorems 7, 8, and 9. First, using part (i) of these theorems, notice that on I_7:

$$C_{Stage\,2}(\bar{Q}^f) = K_1$$

for some constant K_1. (We use $K, K_1, ..., K_8$ to represent constants throughout this proof.) Thus, since $u^f \geq 0$, we have:

$$\bar{Q}_7^f = \underset{\bar{Q}^f \in I_7}{\operatorname{argmin}} \quad u^f \, \bar{Q}^f + C_{Stage\,2}(\bar{Q}^f) = \min_{\bar{Q}^f \in I_7} \bar{Q}^f = \sum_{j=1,2} F_j^{-1}(\frac{p_j + r_j - c_j^f}{p_j + r_j + h_j}).$$

Hence, $\bar{Q}_7^f \in I_6$. Thus, from (5.56), $\bar{Q}^{f*} \in \{\bar{Q}_k^f, k = 1, ..., 6\}$. Next, on I_6, using part (ii) of Theorems 7, 8, and 9: $C_{U,U}(\bar{Q}^f) = C_{U,U}(\bar{Q}^f) = C_{U,D}(\bar{Q}^f) = C_{D,U}(\bar{Q}^f) =$

$C_{D,D}(\bar{Q}^f) = \Gamma(\bar{Q}^f)$. (Notice that although Theorems 7, 8, and 9 are presented based on open end intervals, one can consider closed intervals, since $C_{Stage\,2}(\cdot)$ is continuous at end points.) Therefore,

$$\begin{aligned} C_{Stage\,2}(\bar{Q}^f) &= \theta^1\,\theta^2\,C_{U,U}(\bar{Q}^f) + \theta^1\,(1-\theta^2\,)C_{U,D}(\bar{Q}^f) \\ &\quad +(1-\theta^1)\,\theta^2\,C_{D,U}(\bar{Q}^f) + (1-\theta^1)(1-\theta^2)\,C_{D,D}(\bar{Q}^f) \\ &= \Gamma(\bar{Q}^f), \end{aligned}$$

and

$$\bar{Q}_6^f = \underset{\bar{Q}^f \in I_6}{\operatorname{argmin}}\ u^f\,\bar{Q}^f + C_{Stage\,2}(\bar{Q}^f) = \underset{\bar{Q}^f \in I_6}{\operatorname{argmin}}\ u^f\,\bar{Q}^f + \Gamma(\bar{Q}^f).$$

Next, if $\bar{Q}^f \in I_5$ then, using part (iii) of Theorem 7, $C_{U,U}(\bar{Q}^f) = (c^f - c)\,\bar{Q}^f + K_2$ for some constant K_2. Moreover, from Theorem 8 parts (iii) and (ii) (respectively) we have : $C_{U,D}(\bar{Q}^f) = (c_1^f - c^1)\bar{Q}^f + K_3$, and $C_{D,u} = \Gamma(\bar{Q}^f)$. Also, from Theorem 9 part (iii) $C_{D,D} = \Gamma(\bar{Q}^f)$. Thus,

$$C_{Stage\,2}(\bar{Q}^f) = \big(-\theta^1(c^1 - c_1^f)\big)\bar{Q}^f + \bar{\theta}^1\Gamma(\bar{Q}^f) + K,$$

for some constant K. Hence,

$$\bar{Q}_5^f = \underset{\bar{Q}^f \in I_5}{\operatorname{argmin}}\,\big(u^f - \theta^1(c^1 - c_1^f)\big)\bar{Q}^f + \bar{\theta}^1\Gamma(\bar{Q}^f).$$

Next, if $\bar{Q}^f \in I_4$ then, using part (iii) of Theorems 7 and 8, $C_{U,U}(\bar{Q}^f) = (c^f - c)\,\bar{Q}^f + K_2$, $C_{D,U}(\bar{Q}^f) = (c_2^f - c^2)\bar{Q}^f + K_4$ (for some constants K_2, K_3, K_4). Also, from Theorem 9 part (iii) $C_{D,D} = \Gamma(\bar{Q}^f)$. Hence,

$$\bar{Q}_4^f = \underset{\bar{Q}^f \in I_4}{\operatorname{argmin}}\,\big(u^f - \theta^1(c^1 - c_1^f) - \bar{\theta}^1\theta^2(c^2 - c_2^f)\big)\bar{Q}^f + \bar{\theta}^1\bar{\theta}^2\Gamma(\bar{Q}^f).$$

Similarly, on I_3, from part (iv) of Theorems 7 and 8: $C_{U,U}(\bar{Q}^f) = c_2^f\bar{Q}^f + G_2(\bar{Q}^f) + K_5$,

$C_{U,D}(\bar{Q}^f) = c_2^f\bar{Q}^f + G_2(\bar{Q}^f) + K_6$. Moreover, from part (iii) of Theorem 8, $C_{D,U}(\bar{Q}^f) = (c_2^f - c^2)\bar{Q}^f + K_4$. Also, from Theorem 9 part (ii), $C_{D,D} = \Gamma(\bar{Q}^f)$. Hence,

$$\bar{Q}_3^f = \underset{\bar{Q}^f \in I_3}{\operatorname{argmin}} \left(u^f - \bar{\theta}^1\theta^2(c^2 - c_2^f) + \theta^1 c_2^f\right)\bar{Q}^f + \theta^1 G_2(\bar{Q}^f) + \bar{\theta}^1\bar{\theta}^2\Gamma(\bar{Q}^f).$$

Next, on I_2, from part (v) of Theorem 7, $C_{U,U}(\bar{Q}^f) = (c_2^f - c^2)\bar{Q}^f + K_7$. Also, from Theorem 8, $C_{U,D}(\bar{Q}^f) = c_2^f\bar{Q}^f + G_2(\bar{Q}^f) + K_6$. Moreover, from part (iii) of Theorem 8, $C_{D,U}(\bar{Q}^f) = (c_2^f - c^2)\bar{Q}^f + K_4$. Also, from Theorem 9 part (ii), $C_{D,D} = \Gamma(\bar{Q}^f)$. Thus,

$$\bar{Q}_2^f = \underset{\bar{Q}^f \in I_2}{\operatorname{argmin}} \left(u^f - \theta^2(c^2 - c_2^f) + \theta^1\bar{\theta}^2 c_2^f\right)\bar{Q}^f + \theta^1\bar{\theta}^2 G_2(\bar{Q}^f) + \bar{\theta}^1\bar{\theta}^2\Gamma(\bar{Q}^f).$$

Finally, on I_1, from part (v) of Theorem 7, $C_{U,U}(\bar{Q}^f) = (c_2^f - c^2)\bar{Q}^f + K_7$. Also, from Theorem 8, part (iv), $C_{U,D}(\bar{Q}^f) = c_2^f\bar{Q}^f + G_2(\bar{Q}^f) + K_6$ and $C_{D,U}(\bar{Q}^f) = c_1^f\bar{Q}^f + G_1(\bar{Q}^f) + K_8$. Moreover, from Theorem 9 part (ii), $C_{D,D} = \Gamma(\bar{Q}^f)$. Thus,

$$\begin{aligned}\bar{Q}_1^f \;=\; \underset{\bar{Q}^f \in I_1}{\operatorname{argmin}} \left(u^f - \theta^1\theta^2(c^2 - c_2^f) + \theta^1\bar{\theta}^2 c_2^f + \bar{\theta}^1\theta^2 c_1^f\right)\bar{Q}^f + \bar{\theta}^1\theta^2 G_1(\bar{Q}^f) + \theta^1\bar{\theta}^2 G_2(\bar{Q}^f) \\ + \bar{\theta}^1\bar{\theta}^2\Gamma(\bar{Q}^f),\end{aligned}$$

which completes the proof. □

Proof of Corollary 2: To prove part (i) notice that, from Proposition 10, $\bar{Q}^{f*} \in \{\bar{Q}_k^f, k = 1, ..., 6\}$. Moreover, $\bar{Q}_k^f \in I_k$ (see Proposition 10 for definition of I_k for $k = 1, ..., 6$). Thus, since every member of I_6 is greater than or equal to any member of I_k for $k = 1, ..., 5$, we have:

$$\bar{Q}^{f*} \leq \bar{Q}_6^f \leq \max_{\bar{Q}^f \in I_6} \bar{Q}^f = \sum_{j=1,2} F_j^{-1}\left(\frac{p_j + r_j - c_j^f}{p_j + r_j + h_j}\right).$$

To prove part (ii) notice that, when $\theta^j\,(c^j - c_j^f) \geq u^f$ (for $j = 1, 2$), the optimal capacity reservation level without recourse presented in Theorem 13 is greater than or equal to the upper bound of capacity reservation level with recourse obtained in part (i) (since CDFs $F_j(.)$ are nondecreasing). Hence, when $\theta^j\,(c^j - c_j^f) \geq u^f$ (for $j = 1, 2$), the optimal capacity reservation level with recourse is less than or equal to the optimal capacity reservation level without recourse. □

Proof of Theorem 10: Notice that when parameters are product-independent, each of the intervals I_2, I_3 and I_5 in Proposition 10 only include a single point. Moreover, in that case, $I_2 = I_3 \subset I_4$ and $I_5 \subset I_6$. Thus, by Proposition 10, $\bar{Q}^{f*} \in \{\bar{Q}_1^f, \bar{Q}_4^f, \bar{Q}_6^f\}$. Additionally, using symmetric parameters in Proposition 10, $\Gamma(\bar{Q}^f) = 2\,[c^f\,\bar{Q}^f/2 + G(\bar{Q}^f/2)] = c^f\,\bar{Q}^f + 2\,G(\bar{Q}^f/2)$.

Next notice that since $G(\cdot)$ is a convex function ($G''(\cdot) = (p + r + h)f(\cdot) \geq 0$ using Leibniz rule), $\Gamma(\cdot)$ is a convex function. Thus, $\bar{Q}_1^f, \bar{Q}_4^f$, and $\bar{Q}_6^f$ are minimizers of convex functions on convex (and compact) sets. Thus, we can characterize them using the first order condition. We have:

$$\begin{aligned}\bar{Q}_1^f \;=\; \underset{\bar{Q}^f \in I_1}{\operatorname{argmin}} \big(u^f - \theta^1\theta^2(c^2 - c_2^f) + \theta^1\bar{\theta}^2 c_2^f + \bar{\theta}^1\theta^2 c_1^f\big)\bar{Q}^f + \bar{\theta}^1\theta^2 G_1(\bar{Q}^f) + \theta^1\bar{\theta}^2 G_2(\bar{Q}^f) \\ + \bar{\theta}^1\bar{\theta}^2\Gamma(\bar{Q}^f),\end{aligned}$$

Therefore, to characterize $\bar{Q}_1^f$, using Leibniz rule and setting the derivative of the objective function equal to zero results in the candidate

$$\hat{Q}_1^f = \frac{d\,\big((1 - \theta^1\theta^2)(p + r) + \theta^1\theta^2 c - u^f - c^f\big)}{(\bar{\theta}^1\theta^2 + \theta^1\bar{\theta}^2 + \frac{\bar{\theta}^1\bar{\theta}^2}{2})(p + r + h)}.$$

But $\hat{Q}_1^f \in I_1$ only if C1 does not hold but C2 holds, $\hat{Q}_1^f$ is less than any point in I_1 if C2 does not hold, and $\hat{Q}_1^f$ is greater than any point in I_1 otherwise. Thus,

$$\bar{Q}_1^f = \begin{cases} 0 & : \text{ if C2 does not hold,} \\ \hat{Q}_1^f & : \text{ if C2 holds but C1 does not hold,} \\ d\,(p+r-c)/(p+r+h) & : \text{ otherwise.} \end{cases} \tag{5.57}$$

To characterize $\bar{Q}_4^f$, notice that

$$\bar{Q}_4^f = \underset{\bar{Q}^f \in I_4}{\operatorname{argmin}} \left(u^f - \theta^1(c^1 - c_1^f) - \bar{\theta}^1\theta^2(c^2 - c_2^f)\right)\bar{Q}^f + \bar{\theta}^1\bar{\theta}^2\Gamma(\bar{Q}^f).$$

Thus, using Leibniz rule and setting the derivative of the objective function equal to zero results in the candidate

$$\hat{Q}_4^f = \frac{2d\left(p+r-c-\frac{u^f-(c-c^f)}{\bar{\theta}^1\,\bar{\theta}^2}\right)}{p+r+h}.$$

But $\hat{Q}_4^f$ is in I_4 only if C1 holds and $c - c^f < u^f$, $\hat{Q}_4^f$ is less than any point in I_4 if C1 does not hold and $c - c^f < u^f$, and $\hat{Q}_4^f$ is greater than any point in I_4 if $c - c^f \geq u^f$. Thus,

$$\bar{Q}_4^f = \begin{cases} d\,(p+r-c)/(p+r+h) & : \text{ if C1 does not hold and } c - c^f < u^f, \\ \hat{Q}_4^f & : \text{ if C1 holds and } c - c^f < u^f, \\ 2\,d\,(p+r-c)/(p+r+h) & : \text{ if } c - c^f \geq u^f. \end{cases} \tag{5.58}$$

Similarly, to characterize $\bar{Q}_6^f$, using Leibniz rule and setting the derivative of the objective function equal to zero results in the candidate

$$\hat{Q}_6^f = \frac{2d\left(p+r-c-\frac{u^f-(c-c^f)}{\bar{\theta}^1\,\bar{\theta}^2}\right)}{p+r+h}. \tag{5.59}$$

But $\hat{Q}_6^f \in I_6$ only if $c - c^f \geq u^f$, $\hat{Q}_6^f$ is less than any point in I_4 if $c - c^f < u^f$, and

$\hat{Q}_6^f$ cannot be greater than all points in I_6 (since $u^f \geq 0$). Thus,

$$\bar{Q}_6^f = \begin{cases} 2\,d\,(p+r-u^f-c^f)/(p+r+h) & : \text{ if } c-c^f \geq u^f, \\ 2\,d\,(p+r-c)/(p+r+h) & : \text{ if } c-c^f < u^f. \end{cases} \tag{5.60}$$

Now, since by Proposition 10 $\bar{Q}^{f*} \in \{\bar{Q}_1^f, \bar{Q}_4^f, \bar{Q}_6^f\}$, it remains to compare the cost of $\bar{Q}_1^f, \bar{Q}_4^f$, and $\bar{Q}_6^f$ under different conditions. First, if $c-c^f \geq u^f$, then C1 and C2 trivially hold. Therefore, $\bar{Q}_1^f = d\,(p+r-c)/(p+r+h)$ (from (5.57)), $\bar{Q}_4^f = 2d\,(p+r-c)/(p+r+h)$ (from (5.58)), and $\bar{Q}_6^f = 2d\,(p+r-u^f-c^f)/(p+r+h)$ (from (5.60)). Thus, we notice that $I_1 = [0, \bar{Q}_1^f]$, $I_4 = [\bar{Q}_1^f, \bar{Q}_4^f]$, and $I_6 = [\bar{Q}_4^f, 2\,d\,(p+r-c)/(p+r+h)]$. But since $\bar{Q}_4^f$ is the minimizer over $I_4 = [\bar{Q}_1^f, \bar{Q}_4^f]$, $\bar{Q}_4^f$ has a lower cost than $\bar{Q}_1^f$. Moreover, since $\bar{Q}_6^f$ is the optimizer over $I_6 = [\bar{Q}_4^f, 2\,d\,(p+r-c)/(p+r+h)]$, $\bar{Q}_6^f$ has a lower cost than $\bar{Q}_4^f$. Hence, $\bar{Q}_6^f = 2d\,(p+r-u^f-c^f)/(p+r+h)$ is the optimal solution.

Second, consider the case where $c-c^f < u^f$. If (i) both C1 and C2 do not hold then $\bar{Q}_1^f = 0$, $\bar{Q}_4^f = d\,(p+r-c)/(p+r+h)$, and $\bar{Q}_6^f = 2d\,(p+r-c)/(p+r+h)$. Next notice that $\bar{Q}_4^f$ is a minimizer over $I_4 = [\bar{Q}_4^f, \bar{Q}_6^f]$ and hence has a lower cost than $\bar{Q}_6^f$. Also, $\bar{Q}_1^f = 0$ is a minimizer over $I_1 = [0, \bar{Q}_4^f]$ and hence has a lower cost than $\bar{Q}_4$. Thus, in this case, $\bar{Q}^{f*} = \bar{Q}_1^f = 0$. However, if (ii) C1 does not hold but C2 holds then $\bar{Q}_1^f = \hat{Q}_1^f$, $\bar{Q}_4^f = d\,(p+r-c)/(p+r+h)$, and $\bar{Q}_6^f = 2d\,(p+r-c)/(p+r+h)$. Next notice that $\bar{Q}_4^f$ is a minimizer over $I_4 = [\bar{Q}_4^f, \bar{Q}_6^f]$ and hence has a lower cost than $\bar{Q}_6^f$. Also, $\bar{Q}_1^f = 0$ is a minimizer over $I_1 = [0, \bar{Q}_4^f]$ and hence has a lower cost than $\bar{Q}_4$. Thus, in this case, $\bar{Q}^{f*} = \hat{Q}_1^f$.

Next if (iii) C1 holds, first consider the case where C2 holds as well. In this case, $\bar{Q}_1^f = d\,(p+r-c)/(p+r+h)$, $\bar{Q}_4^f = \hat{Q}_4^f$ and $\bar{Q}_6^f = 2d\,(p+r-c)/(p+r+h)$. Next since $\bar{Q}_4^f = \hat{Q}_4^f$ is a minimizer over $I_4 = [\bar{Q}_1^f, \bar{Q}_6^f]$, $\bar{Q}_4^f = \hat{Q}_4^f$ has lower cost than $\bar{Q}_1^f$ and $\bar{Q}_6^f$. Thus in this case $\bar{Q}^{f*} = \hat{Q}_4^f$. To complete part (iii) now suppose C1 holds but C2 does not. Then $\bar{Q}_1^f = 0$, $\bar{Q}_4^f = \hat{Q}_4^f$, and $\bar{Q}_6^f = 2d\,(p+r-c)/(p+r+h)$. Next, similar to the previous case, since $\bar{Q}_6^f \in I_4$ and $\bar{Q}_4^f$ is a minimizer over I_4, $\bar{Q}_4^f = \hat{Q}_4^f$ has a

lower cost than $\bar{Q}_6^f$. Therefore, $\bar{Q}^{f*} \in \{0, \hat{Q}_4^f\}$. To determine Q^{f*} in this case, hence, it is sufficient to compare the cost of options $\bar{Q}^f = 0$ and $\bar{Q}^f = \hat{Q}_4^f$. To compute $C_{Stage2}(\bar{Q}_1^f = 0)$, using part (v) of Theorem 7, part (iv) of Theorem 8, and part (ii) of Theorem 9 we have:

$$\begin{aligned}
C_{U,U}(0) &= \frac{2\,c\,d(p+r-c)}{p+r+h} + 2G(\frac{d(p+r-c)}{p+r+h}),\\
C_{U,D}(0) &= \frac{c\,d(p+r-c)}{p+r+h} + G(\frac{d(p+r-c)}{p+r+h}) + G(0),\\
C_{D,U}(0) &= \frac{c\,d(p+r-c)}{p+r+h} + G(\frac{d(p+r-c)}{p+r+h}) + G(0),\\
C_{D,D}(0) &= 2\,G(0).
\end{aligned}$$

Hence,

$$\begin{aligned}
C_{Stage\,2}(0) &= \theta^1\,\theta^2\,C_{U,U}(0) + \theta^1\,(1-\theta^2\,)C_{U,D}(0)\\
&+ (1-\theta^1)\,\theta^2\,C_{D,U}(0) + (1-\theta^1)(1-\theta^2)\,C_{D,D}(0)\\
&= (\theta^1\bar{\theta}^2 + \bar{\theta}^1\theta^2 + 2\theta^1\theta^2)\big[\frac{c\,d(p+r-c)}{p+r+h} + G(\frac{d(p+r-c)}{p+r+h})\big]\\
&+ (\theta^1\bar{\theta}^2 + \bar{\theta}^1\theta^2 + 2\bar{\theta}^1\bar{\theta}^2)G(0).
\end{aligned}$$

Moreover, since $\bar{Q}^f = 0$, $u^f\bar{Q}^f + C_{Stage\,2}(0) = C_{Stage\,2}(0)$. Next, we compute cost of the option $\bar{Q}^f = \hat{Q}_4^f$, and compare it with the cost of $\bar{Q}^f = 0$ computed above. Since $\hat{Q}_4^f \in I_4$, we shall use part (iii) of using Theorems 7, 8, and part (i) of Theorem 9. Doing so we have:

$$
\begin{aligned}
C_{U,U}(\hat{Q}_4^f) &= (c^f - c)\hat{Q}_4^f + 2[\frac{c\,d(p+r-c)}{p+r+h} + G(\frac{d(p+r-c)}{p+r+h})],\\
C_{U,D}(\hat{Q}_4^f) &= (c^f - c)\hat{Q}_4^f + 2[\frac{c\,d(p+r-c)}{p+r+h} + G(\frac{d(p+r-c)}{p+r+h})],\\
C_{D,U}(\hat{Q}_4^f) &= (c^f - c)\hat{Q}_4^f + 2[\frac{c\,d(p+r-c)}{p+r+h} + G(\frac{d(p+r-c)}{p+r+h})],\\
C_{D,D}(\hat{Q}_4^f) &= c^f\,\hat{Q}_4^f + 2G(\hat{Q}_4^f/2).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
C_{Stage\,2}(\hat{Q}_4^f) &= \theta^1\,\theta^2\,C_{U,U}(\hat{Q}_4^f) + \theta^1\,(1-\theta^2\,)C_{U,D}(\hat{Q}_4^f)\\
&+ (1-\theta^1)\,\theta^2\,C_{D,U}(\hat{Q}_4^f) + (1-\theta^1)(1-\theta^2)\,C_{D,D}(\hat{Q}_4^f)\\
&= [c^f - (1-\bar{\theta}^1\bar{\theta}^2)c]\hat{Q}_4^f + 2(1-\bar{\theta}^1\bar{\theta}^2)[\frac{c\,d(p+r-c)}{p+r+h} + G(\frac{d(p+r-c)}{p+r+h})]\\
&+ 2\bar{\theta}^1\bar{\theta}^2 G(\hat{Q}_4^f/2),
\end{aligned}
$$

and the total optimal cost with $\bar{Q}^f = \hat{Q}_4^f$ is $u^f\hat{Q}_4^f + C_{Stage\,2}(\hat{Q}_4^f)$. Thus, denoting the optimal cost with $\bar{Q}^f = \hat{Q}_4^f$ minus the optimal cost with $\bar{Q}^f = 0$ by Δ, (after simplification) we have :

$$
\begin{aligned}
\Delta &= \big((1-\theta^1\theta^2) + \bar{\theta}^1\bar{\theta}^2\big)\big[\frac{c\,d(p+r-c)}{p+r+h} + G(\frac{d(p+r-c)}{p+r+h}) - G(0)\big] \quad (5.61)\\
&+ \big(u^f + c^f - (1-\bar{\theta}^1\bar{\theta}^2)c\big)\hat{Q}_4^f + 2\bar{\theta}^1\bar{\theta}^2\big[G(\hat{Q}_4^f/2) - G(0)\big]. \quad (5.62)
\end{aligned}
$$

Next notice that both (5.61) and (5.62) are nonpositive. To see that (5.61) is nonpositive, define $g(x) = cx - G(x) - G(0)$ and notice that (5.61) is equal to $\big((1-\theta^1\theta^2) + \bar{\theta}^1\bar{\theta}^2\big)g(q^*)$ with $q^* = F^{-1}(\frac{p+r-c}{p+r+h}) = \frac{d(p+r-c)}{p+r+h}$. Hence, (5.61) is nonpositive, since $g(0) = 0$ and q^* is the minimizer of $g(\cdot)$ ($G(\cdot)$ is convex and q^* is the solution to the first order condition). Next, to see that (5.62) is nonpositive, define

$$
\hat{g}(x) = \big(u^f + c^f - (1-\bar{\theta}^1\bar{\theta}^2)c\big)x + 2\bar{\theta}^1\bar{\theta}^2\big[G(x/2) - G(0)\big],
$$

and notice that (5.62) is equal to $\hat{g}(\hat{Q}_4^f)$. Thus, (5.62) is nonpositive, since $\hat{g}(0) = 0$ and $\hat{g}(\cdot)$ is a convex function minimized at $\hat{Q}_4^f$ (see the definition of $\hat{Q}_4^f$ or check the first and second order conditions of $\hat{g}(\cdot)$). Hence, $\Delta \leq 0$. Thus, $\bar{Q}^{f*} = \hat{Q}_4^f$, and the proof is complete. □

Proof of Theorem 11: Notice that program (5.12) is convex with linear constraints, and therefore KKT conditions are necessary and sufficient to characterize the optimal solution. Assume λ_j^f $(j = 1, 2)$ and μ represent the Lagrangian multipliers corresponding to constraints $q_j^f \geq 0$ and $q_1^f + q_2^f \leq \bar{Q}^f$, respectively. Using the Leibniz rule, KKT conditions are:

$$
\begin{aligned}
& q_1^f + q_2^f \leq \bar{Q}^f \\
& c_j^f + (h_j + p_j + r_j)F_j(q^j + q_j^f) - p_j - r_j = \lambda_j^f - \mu \qquad (j = 1, 2) \qquad (5.63) \\
& \mu(q_1^f + q_2^f - \bar{Q}^f) = 0 \\
& \lambda_j^f q_j^f = 0 \qquad (j = 1, 2) \\
& q_j^f, \mu, \lambda_j^f \geq 0. \qquad (j = 1, 2)
\end{aligned}
$$

Condition (5.63) results in:

$$
q_j^f = F_j^{-1}\left(\frac{p_j + r_j - c_j^f + \lambda_j^f - \mu}{p_j + r_j + h_j}\right) - q^j. \qquad (j = 1, 2)
$$

Hence, the KKT conditions can be written as follows:

$$q_1^f + q_2^f \leq \bar{Q}^f \tag{5.64}$$

$$q_j^f = F_j^{-1}(\frac{p_j + r_j - c_j^f + \lambda_j^f - \mu}{p_j + r_j + h_j}) - q^j \qquad (j = 1, 2) \tag{5.65}$$

$$\mu(q_1^f + q_2^f - \bar{Q}^f) = 0 \tag{5.66}$$

$$\lambda_j^f q_j^f = 0 \qquad (j = 1, 2) \tag{5.67}$$

$$q_j^f, \mu, \lambda_j^f \geq 0. \qquad (j = 1, 2) \tag{5.68}$$

Now, it is trivial to show that the optimal solutions α_j and β_j satisfy conditions (5.64) - (5.68) in the appropriate range of $\bar{Q}^f$ defined in the theorem. □

Proof of Theorem 12: The proof follows from Theorem 11 after setting $q^n = 0$. □

Proof of Theorem 13:

Let μ, λ^j and λ_j^f for $(j = 1, 2)$ denote the Lagrangian multipliers, respectively, for constraints (5.94)-(5.95). KKT conditions are then:

$$q_1^f + q_2^f \leq \bar{Q}^f \tag{5.69}$$

$$\partial C_P/\partial q^j - \lambda^j = 0 \qquad (j = 1, 2) \tag{5.70}$$

$$\partial C_P/\partial q_j^f - \lambda_j^f + \mu = 0 \qquad (j = 1, 2) \tag{5.71}$$

$$\partial C_P/\partial \bar{Q}^f - \mu = 0 \qquad (j = 1, 2) \tag{5.72}$$

$$\mu(q_1^f + q_2^f - \bar{Q}^f) = 0 \tag{5.73}$$

$$\lambda^j q^j = 0 \qquad (j = 1, 2) \tag{5.74}$$

$$\lambda_j^f q_j^f = 0 \qquad (j = 1, 2) \tag{5.75}$$

$$q^j, q_j^f, \mu, \lambda^j, \lambda_j^f \geq 0. \qquad (j = 1, 2) \tag{5.76}$$

Moreover, using Leibniz rule we have:

$$\partial C_P/\partial q^j = \theta^j[c^j + (h_j + p_j + r_j)F_j(q^j + q_j^f) - p_j - r_j] \qquad (j = 1, 2)$$

$$\partial C_P/\partial q_j^f = c_j^f + \theta^j[(h_j + p_j + r_j)F_j(q^j + q_j^f) - p_j - r_j] + (1 - \theta^j)[(h_j + p_j + r_j)F_j(q_j^f) - p_j - r_j] \qquad (j = 1, 2)$$

$$\partial C_P/\partial \bar{Q}^f = u_j^f. \qquad (j = 1, 2)$$

Hence, conditions (5.69)-(5.76) are:

$$q_1^f + q_2^f \leq \bar{Q}^f$$

$$\theta^j[c^j + (h_j + p_j + r_j)F_j(q^j + q_j^f) - p_j - r_j] = \lambda^j \qquad (j = 1, 2)$$

$$c_j^f + \theta^j[(h_j + p_j + r_j)F_j(q^j + q_j^f) - p_j - r_j] + (1 - \theta^j)[(h_j + p_j + r_j)F_j(q_j^f) - p_j - r_j] = \lambda_j^f - \mu \qquad (j = 1, 2)$$

$$\mu = u_j^f \qquad (j = 1, 2)$$

$$\mu(q_1^f + q_2^f - \bar{Q}^f) = 0$$

$$\lambda^j q^j = 0 \qquad (j = 1, 2)$$

$$\lambda_j^f q_j^f = 0 \qquad (j = 1, 2)$$

$$q^j, q_j^f, \mu, \lambda^j, \lambda_j^f \geq 0, \qquad (j = 1, 2)$$

or equivalently:

$$F_j(q^{j*} + q_j^{f*}) = \frac{\frac{\lambda^j}{\theta^j} + p_j + r_j - c^j}{p_j + r_j + h_j} \qquad (j = 1, 2) \qquad (5.77)$$

$$F_j(q_j^{f*}) = \frac{\frac{\lambda_j^f - \lambda^j + \theta^j c^j - u_j^f}{1-\theta^j} + p_j + r_j}{p_j + r_j + h_j} \qquad (j = 1, 2) \qquad (5.78)$$

$$\bar{Q}f^* = q_1^{f*} + q_2^{f*}$$

$$\lambda^j q^{j*} = 0 \qquad (j = 1, 2) \qquad (5.79)$$

$$\lambda_j^f q_j^{f*} = 0 \qquad (j = 1, 2) \qquad (5.80)$$

$$q^j, q_j^f, \lambda^j, \lambda_j^f \geq 0. \qquad (j = 1, 2) \qquad (5.81)$$

Since by assumption $u_j^f > c^j$ and $h_j \geq 0$, we have $\frac{\frac{\theta^j c^j - u_j^f}{1-\theta^j} + p_j + r_j}{p_j + r_j + h_j} \leq 1$. Hence, if $\frac{\frac{\theta^j c^j - u_j^f}{1-\theta^j} + p_j + r_j}{p_j + r_j + h_j} \geq 0$, setting $\lambda^j = \lambda_j^f = 0$ (for every $j \in \{1, 2\}$ that satisfies this inequality) to guarantee conditions (5.79)-(5.81) results in the optimal solution:

$$q_j^{f*} = F_j^{-1}\left(\frac{p_j + r_j - \frac{u_j^f - \theta^j c^j}{1-\theta^j}}{p_j + r_j + h_j}\right) \qquad (j = 1, 2)$$

$$q^{j*} = F_j^{-1}\left(\frac{p_j + r_j - c^j}{p_j + r_j + h_j}\right) - F_j^{-1}\left(\frac{p_j + r_j - \frac{u_j^f - \theta^j c^j}{1-\theta^j}}{p_j + r_j + h_j}\right) \qquad (j = 1, 2)$$

$$\bar{Q}f^* = \sum_{j=1}^{2} q_j^{f*} = \sum_{j=1}^{2} F_j^{-1}\left(\frac{p_j + r_j - \frac{u_j^f - \theta^j c^j}{1-\theta^j}}{p_j + r_j + h_j}\right).$$

However, if $\frac{\frac{\theta^j c^j - u_j^f}{1-\theta^j} + p_j + r_j}{p_j + r_j + h_j} < 0$, Eq. (5.78) can not be satisfied with $\lambda^j = \lambda_j^f = 0$. In this case, however, setting $q_j^{f*} = \lambda^j = 0$ satisfies all KKT conditions and hence is

optimal. Therefore, because $F_j(0) = 0$, a general optimal solution for $0 \leq \theta^j < 1$ is:

$$q_j^{f*} = F_j^{-1}([\frac{p_j + r_j - \frac{u_j^f - \theta^j c^j}{1-\theta^j}}{p_j + r_j + h_j}]^+) \qquad (j = 1, 2)$$

$$q^{j*} = F_j^{-1}(\frac{p_j + r_j - c^j}{p_j + r_j + h_j}) - F_j^{-1}([\frac{p_j + r_j - \frac{u_j^f - \theta^j c^j}{1-\theta^j}}{p_j + r_j + h_j}]^+) \qquad (j = 1, 2)$$

$$\bar{Q^f}^* = \sum_{j=1}^{2} q_j^{f*} = \sum_{j=1}^{2} F_j^{-1}([\frac{p_j + r_j - \frac{u_j^f - \theta^j c^j}{1-\theta^j}}{p_j + r_j + h_j}]^+)$$

and the proof is complete. □

Proof of Lemma 5: To obtain optimal decisions in the absence of the flexible supplier, set $q_1^f = q_2^f = \bar{Q}^f = 0$ in (5.92) and use Leibniz rule to derive first order condition (or simply use the results of a basic newsvendor model with lost sales). Doing that we get $q'^{j*} = F_j^{-1}(\frac{p_j + r_j - c^j}{p_j + r_j + h_j})$. Then using Eq. (5.96) we have

$$C_T\left(q'^{1*}, q'^{2*}, 0, 0, 0\right) =$$

$$\sum_{j=1}^{2}[(1-\pi_0^j)(c^j F_j^{-1}(\frac{p_j + r_j - c^j}{p_j + r_j + h_j}) + G_j(F_j^{-1}(\frac{p_j + r_j - c^j}{p_j + r_j + h_j}))) + \pi_0^j G_j(0)].$$

Also, to obtain $C_T\left(q^{1*}, q^{2*}, q_1^{f*}, q_2^{f*}, \bar{Q}^{f*}\right)$, substitute the perceived optimal values (derived by Theorem 13) in Eq. (5.96). Hence, using Eq. (5.21) and after simplification we have

$$\begin{aligned} V^f &= C_T\left(q'^{1*}, q'^{2*}, 0, 0, 0\right) - C_T\left(q^{1*}, q^{2*}, q_1^{f*}, q_2^{f*}, \bar{Q}^{f*}\right) \\ &= \sum_{j=1}^{2}[\pi_0^j\left(G_j(0) - G_j(q^{f*})\right) - (u_j^f - (1-\pi_0^j)c^j)q_j^{f*}]. \end{aligned}$$

Now, replacing $G_j(0) = p_j E(D_j)$ in above equation completes the proof. □

Proof of Theorem 14: To prove part (i), note that the value of the flexible supplier as perceived by the firm (and not its true value) is

$$V^f = \sum_{j=1}^{2} (1-\theta^j)(p_j E(D_j) - G_j(q_j^{f*})) - (u_j^f - \theta^j c^j) q_j^{f*},$$

where for simplicity we have removed subscript P to denote that V^f here is a perceived value. We first show that V^f as perceived by the firm is positive if and only if

$$\exists j \in \{1,2\} : \; q_j^{f*} = F_j^{-1}([\frac{p_j + r_j - \frac{u_j^f - \theta^j c^j}{1-\theta^j}}{p_j + r_j + h_j}]^+) > 0.$$

To show this, for $x \in [0,\infty)$ let $V_j^f(x) = (1-\theta^j)(p_j E(D_j) - G_j(x)) - (u_j^f - \theta^j c^j)\, x$ denote the perceived value of the flexible supplier with respect to product j if the firm orders x units of product j from the flexible supplier. Then we have

$$\frac{\partial(V_j^f)}{\partial x} = -(1-\theta^j)((h_j + p_j + r_j)F_j(x) - p_j - r_j) + \theta^j c^j - u_j^f.$$

Hence, $\frac{\partial}{\partial x}(\frac{\partial(V_j^f)}{\partial x}) = -(1-\theta^j)\,(h_j + p_j + r_j) f_j(x) \leq 0$. Therefore, $V_j^f(\cdot)$ is concave and first order condition yields

$$\max\{V_j^f(x) : x \in [0,\infty)\} = V_j^f(q_j^{f*})\ .$$

Additionally, because $G_j(0) = p_j E(D_j)$, we get $V_j^f(0) = 0$. Hence, we have $V_j^f(q_j^{f*}) \geq 0$. Also, $\frac{\partial(V_j^f)}{\partial x} > 0$ for $x \in [0, q_j^{f*})$ shows that $V_j^f(\cdot)$ is increasing in $[0, q_j^{f*})$. Hence, $V_j^f(q_j^{f*}) > 0$ if, and only if, $q_j^{f*} > 0$. Now, note that the value of the flexible supplier as perceived by the firm is $V^f = \sum_{j=1}^{2} V_j^f(q_j^{f*})$. Therefore, $V^f > 0$ if, and only if,

$$\exists j \in \{1,2\} : \; q_j^{f*} = F_j^{-1}([\frac{p_j + r_j - \frac{u_j^f - \theta^j c^j}{1-\theta^j}}{p_j + r_j + h_j}]^+) > 0.$$

Moreover, because $F_j(\cdot)$ is non-decreasing and $F_j(x) > 0$ for all $x > 0$, the above condition is equivalent to

$$\exists j \in \{1,2\} : \; p_j + r_j - \frac{u_j^f - \theta^j c^j}{1-\theta^j} > 0,$$

which is equivalent to

$$\exists j \in \{1,2\} : (1-\theta^j)(p_j + r_j) > u_j^f - \theta^j c^j,$$

or similarly

$$\exists j \in \{1,2\} : p_j + r_j - u_j^f > \theta^j (p_j + r_j - c^j).$$

This is equivalent to (5.23) (as $r_j > c^j$ and hence $(p_j + r_j - c^j) > 0$).

We prove part (ii) by contradiction. First, note that by Lemma 5, the true value of the flexible supplier for the firm is

$$V^f = \sum_{j=1}^{2} [\pi_0^j \left(p_j E(D_j) - G_j(q_j^{f*})\right) - (u_j^f - (1-\pi_0^j)c^j) q_j^{f*}].$$

Now suppose $V^f > 0$. Because $q_1^{f*} = q_2^{f*} = 0$ results in $V^f = 0$, we then have: $\exists j \in \{1,2\} : \; q_j^{f*} > 0$ or equivalently (by proof of part (i)): $\exists j \in \{1,2\} : \; \theta^j < \frac{p_j + r_j - u_j^f}{p_j + r_j - c^j}$. Since the proof for other cases can be obtained in the same way (merely a change of notation), without loss of generality suppose this is true for j=1, i.e., we have

$$\theta^1 < \frac{p_1 + r_1 - u_1^f}{p_1 + r_1 - c^1} \tag{5.82}$$

and $q_2^{f*} = 0$. We then show that it yields

$$\max\{\theta^1, 1-\pi_0^1\} < \frac{p_1 + r_1 - u_1^f}{p_1 + r_1 - c^1}, \tag{5.83}$$

which is a contradiction with the condition given in (5.24) that for both $j = 1, 2$: $\max\{\theta^j, 1-\pi_0^j\} \geq \frac{p_j+r_j-u_j^f}{p_j+r_j-c^j}$. To show that we get (5.83), note that using (5.82), we just need to show that we get

$$1 - \pi_0^1 < \frac{p_1 + r_1 - u_j^f}{p_1 + r_1 - c^1} \tag{5.84}$$

for the case where $\theta^1 < 1 - \pi_0^1$. To show that we get (5.84) in this case, let $V_j^f(x) = \pi_0^j(p_j E(D_j) - G_j(x)) - (u_j^f - (1-\pi_0^j)c^j)\, x$ for $x \in [0, \infty)$ denote the true value of the flexible supplier with respect to product j if the firm orders x units of product j from the flexible supplier. Now, consider the fact that

$$\frac{\partial(V_j^f)}{\partial x} = -\pi_0^j((h_j + p_j + r_j)F_j(x) - p_j - r_j) + (1-\pi_0^j)c^j - u_j^f$$

and $\frac{\partial}{\partial x}(\frac{\partial(V_j^f)}{\partial x}) = -\pi_0^j\,(h_j + p_j + r_j)f_j(x) \leq 0$. Hence, $V_j^f(\cdot)$ defined above is concave and the first order condition yields

$$\max\{V_j^f(x) : x \in [0,\infty)\} = V_j^f(F_j^{-1}(\frac{p_j + r_j - \frac{u_j^f - (1-\pi_0^j)c^j}{\pi_0^j}}{p_j + r_j + h_j})).$$

Now, since $\theta^1 < 1 - \pi_0^1$, we have

$$\frac{p_j + r_j - \frac{u_j^f - (1-\pi_0^j)c^j}{\pi_0^j}}{p_j + r_j + h_j} < \frac{p_j + r_j - \frac{u_j^f - \theta^j c^j}{1-\theta^j}}{p_j + r_j + h_j} \leq F_j(q_j^{f*}).$$

Therefore, as $F_j(\cdot)$ is non-decreasing, we have $F_j^{-1}(\frac{p_j+r_j-\frac{u_j^f-(1-\pi_0^j)c^j}{\pi_0^j}}{p_j+r_j+h_j}) \leq q_j^{f*}$. Furthermore, because $\frac{\partial(V_j^f)}{\partial x} < 0$ in the interval $(F_j^{-1}(\frac{p_j+r_j-\frac{u_j^f-(1-\pi_0^j)c^j}{\pi_0^j}}{p_j+r_j+h_j}), +\infty)$, $V_j^f(x)$ is strictly decreasing in this interval. Hence, considering the initial assumptions that

$V^f > 0$ and $q_2^{f*} = 0$ we have

$$0 < V^f = V_1^f(q_1^{f*}) + V_2^f(q_2^{f*}) < V_1^f(F_j^{-1}(\frac{p_j + r_j - \frac{u_j^f - (1-\pi_0^j)c^j}{\pi_0^j}}{p_j + r_j + h_j})) + 0.$$

Now because $V_1^f(0) = 0$ and $F_j(\cdot)$ is non-decreasing and $F_j(x) > 0$ for all $x > 0$, the above condition yields

$$\frac{p_j + r_j - \frac{u_j^f - (1-\pi_0^j)c^j}{\pi_0^j}}{p_j + r_j + h_j} > 0,$$

which is equivalent to (5.84), which in turn implies (5.83). But this is a contradiction and hence the proof is complete. □

Proof of Proposition 11: It is sufficient to show that for both $j = 1, 2$: $\frac{\partial(V^f)}{\partial \epsilon^j} \leq 0$ if $\epsilon^j > 0$ and $\frac{\partial(V^f)}{\partial \epsilon^j} \geq 0$ otherwise. To show this note that by Lemma (5) we have

$$V^f(\epsilon^1, \epsilon^2) = \sum_{j=1}^{2} [\pi_0^j \left(p_j E(D_j) - G_j(q_j^{f*})\right) - (u_j^f - (1 - \pi_0^j)c^j)q_j^{f*}]$$

where using Theorem (13),

$$q_j^{f*} = F_j^{-1}([\frac{p_j + r_j - \frac{u_j^f - (1-\pi_0^j+\epsilon^j)c^j}{\pi_0 - \epsilon^j}}{p_j + r_j + h_j}]^+). \tag{5.85}$$

Hence, using the technique for computing derivative of an inverse function, we have

$$\frac{\partial(q_j^{f*})}{\partial \epsilon^j} = -\frac{(u_j^f - c^j)}{(p_j + r_j + h_j)(\pi_0 - \epsilon^j)^2 f_j(q_j^{f*})} \leq 0 \tag{5.86}$$

since $u_j^f \geq c^j$ for both $j = 1, 2$. Moreover, $\frac{\partial(V^f)}{\partial \epsilon^j} = \frac{\partial(V^f)}{\partial q_j^{f*}} \frac{\partial q_j^{f*}}{\partial \epsilon^j}$. Therefore, $\frac{\partial(V^f)}{\partial \epsilon^j} =$

$$-\frac{(u_j^f - c^j)}{(p_j + r_j + h_j)(\pi_0 - \epsilon^j)^2 f_j(q_j^{f*})}[-\pi_0^j((p_j+r_j+h_j)F_j(q_j^{f*})-p_j-r_j)-(u_j^f-(1-\pi_0)c^j)]. \tag{5.87}$$

Now, notice that because of (5.85), $-\pi_0^j((p_j+r_j+h_j)F_j(q_j^{f*})-p_j-r_j)-(u_j^f-(1-\pi_0)c^j)$ is not negative if $\epsilon^j \geq 0$ and is not positive if $\epsilon^j \leq 0$. Hence, using inequality (5.86) and Eq. (5.87), the proof is complete. □

Proof of Lemma 6: If the firm knows that dedicated supplier j is up, it would not reserve the flexible supplier's capacity for product j and would just order from dedicated supplier j because this supplier is cheaper than the flexible one. The optimal ordering quantity of this product in this case can be determined using a simple single source newsvendor problem. This quantity is the same as total optimal ordering quantity for product j in the general no-information case, i.e., $q_j^{f*}+q^{j*}$ where q^{f*} and q^{j*} are presented in Theorem 13. Using these optimal quantities we would have:

$$\begin{aligned} V^{+j} &= C_T\left(q^{1*}, q^{2*}, q_1^{f*}, q_2^{f*}, \bar{Q}^{f*}\right) - C_T\left(q^{1\#}, q^{2\#}, q_1^{f\#}, q_2^{f\#}, \bar{Q}^{f\#} | i^j > 0\right) \\ &= u_j^f q_j^{f*} + \left[\,(1-\pi_0^j)(c^j q^{j*} + G_j(q^{j*} + q_j^{f*})) + \pi_0^j\, G_j(q_j^{f*})\,\right] \\ &\quad - G_j(q_j^{f*} + q^{j*}) - c^j(q_j^{f*} + q^{j*}). \end{aligned}$$

Then, simplification results in Eq. (5.27). To derive V^{0j}, note that if the firm knows that dedicated supplier j is in the down state (i.e., is disrupted), it procures only from the flexible supplier (for product j) and will solve a single source newsvendor problem with procurement cost of u_j^f. Hence, in that case, its optimal cost for this product is: $G_j(F_j^{-1}(\frac{p_j+r_j-u_j^f}{p_j+r_j+h_j})) + u_j^f F_j^{-1}(\frac{p_j+r_j-u_j^f}{p_j+r_j+h_j})$. Then, using Eq. (5.25) and simplification yields Eq. (5.28). Moreover, by implementing these values for V^{+j} and V^{0j}, the value of information on the threat level of dedicated supplier j (VI^j) can be obtained using

its definition (i.e., $VI^j = (1-\pi_0^j)\,V^{+j} + \pi_0^j\,V^{0j}$). Using this and simplification then results in Eq. (5.29) and the proof is complete. □

Proof of Proposition 12: From Lemmas 5 and 6 (and after simplification) we have:

$$\begin{aligned} VI - V^f &= \sum_{j=1}^{2} \Big[2(u_j^f - c^j) q_j^{f*} + \pi_0^j \,[2\left(G_j(q_j^{f*}) + c^j q_j^{f*}\right) - \big(G_j(F_j^{-1}(\frac{p_j + r_j - u_j^f}{p_j + r_j + h_j})) \\ &+ u_j^f F_j^{-1}(\frac{p_j + r_j - u_j^f}{p_j + r_j + h_j}) + p_j E(D_j)\big)]\Big]. \end{aligned}$$

Next if

$$2\left(G_j(q_j^{f*}) + c^j q_j^{f*}\right) - \big(G_j(F_j^{-1}(\frac{p_j + r_j - u_j^f}{p_j + r_j + h_j})) + u_j^f F_j^{-1}(\frac{p_j + r_j - u_j^f}{p_j + r_j + h_j}) + p_j E(D_j)\big) \geq 0,$$

let $\hat{\pi}_0^j = 1$. Otherwise, let $\hat{\pi}_0^j =$

$$\min\Big\{ \frac{-2\left(u_j^f - c^j\right) q_j^{f*}}{2\left(G_j(q_j^{f*}) + c^j q_j^{f*}\right) - \big(G_j(F_j^{-1}(\frac{p_j + r_j - u_j^f}{p_j + r_j + h_j})) + u_j^f F_j^{-1}(\frac{p_j + r_j - u_j^f}{p_j + r_j + h_j}) + p_j E(D_j)\big)}, 1 \Big\},$$

and notice that since $u_j^f \geq c^j$, $VI - V^f \geq 0$ whenever $\pi_0^j \leq \hat{\pi}_0^j$.

Proof of Proposition 13: To show the first part, it is sufficient to show that for both $j = 1, 2$: $\frac{\partial(VI)}{\partial \epsilon^j} \leq 0$ if $\epsilon^j < 0$ and $\frac{\partial(VI)}{\partial \epsilon^j} \geq 0$ if $\epsilon^j > 0$. To show this note that $\frac{\partial(VI)}{\partial \epsilon^j} = \frac{\partial(VI)}{\partial q_j^{f*}} \frac{\partial q_j^{f*}}{\partial \epsilon^j}$. Moreover, recall that:

$$q_j^{f*} = F_j^{-1}([\frac{p_j + r_j - \frac{u_j^f - (1 - \pi_0^j + \epsilon^j)c^j}{\pi_0 - \epsilon^j}}{p_j + r_j + h_j}]^+). \tag{5.88}$$

Hence, using the technique to get the derivative of an inverse function we have

$$\frac{\partial(q_j^{f*})}{\partial \epsilon^j} = -\frac{(u_j^f - c^j)}{(p_j + r_j + h_j)(\pi_0 - \epsilon^j)^2 f_j(q_j^{f*})} \leq 0, \tag{5.89}$$

since $u_j^f \geq c^j$ for both $j = 1, 2$. Therefore, to show the result, it is sufficient to show that $\frac{\partial(VI)}{\partial q_j^{f*}} \leq 0$ if $\epsilon^j > 0$ and $\frac{\partial(VI)}{\partial q_j^{f*}} \geq 0$ if $\epsilon^j < 0$. But from Lemma 3 we have

$$VI^j = (u_j^f - (1-\pi_0^j)\, c^j) q_j^{f*} + \pi_0^j \left[G_j(q_j^{f*}) - G_j(F_j^{-1}(\frac{p_j + r_j - u_j^f}{p_j + r_j + h_j})) - u_j^f F_j^{-1}(\frac{p_j + r_j - u_j^f}{p_j + r_j + h_j}) \right] \tag{5.90}$$

and $VI = \sum_{j=1}^{2} VI^j$. Additionally, using Leibniz rule and Eq. (5.2) we get

$$\frac{\partial G_j(q^{f*})}{\partial q_j^{f*}} = [(h_j + p_j + r_j) F_j(q_j^{f*}) - p_j - r_j].$$

Hence, using Eq. (5.90) and replacing q_j^{f*} from Eq. (5.88) in addition to replacing π_0^j by $1 - \theta^j + \epsilon^j$ yields

$$\frac{\partial(VI)}{\partial q_j^{f*}} = -\frac{\epsilon^j (u_j^f - c^j)}{1 - \theta^j}.$$

Therefore, since $u_j^f \geq c^j$ and $\theta^j < 1$ for both $j = 1, 2$, we have $\frac{\partial(VI)}{\partial q_j^{f*}} \leq 0$ if $\epsilon^j > 0$ and $\frac{\partial(VI)}{\partial q_j^{f*}} \geq 0$ if $\epsilon^j < 0$ for both $j = 1, 2$. Hence, the proof of the first part is complete. To show the second part (i.e., to show that $VI(\cdot, \cdot)$ is non-decreasing in $y^j = |\epsilon^j|$ for both $j = 1, 2$), notice that $\frac{\partial(VI)}{\partial y^j} = \frac{\partial(VI)}{\partial \epsilon^j} \frac{\partial \epsilon^j}{\partial y^j}$. Thus, to show $\frac{\partial(VI)}{\partial y^j} \geq 0$, it is sufficient to show that (for both $j = 1, 2$) $\frac{\partial(VI)}{\partial \epsilon^j} \leq 0$ if $\epsilon^j < 0$ and $\frac{\partial(VI)}{\partial \epsilon^j} \geq 0$ if $\epsilon^j > 0$, which is shown above. □

5.9 Appendix B: Parameter Settings

The different parameter settings considered to illustrate the different behaviors, results, and insights in the two Studies of Appendix C are as follow. In Table 5.4, setting 2 represent a case with much higher marginal revenues for the products than setting 1. Setting 3 includes changes in other parameters which result in different critical ratios.

Table 5.4: Suite of Parameter Settings in Studies 4 and 5.

Setting No.	$u_1^f = u_2^f$	j	p_j	r_j	h_j	π_0^j	c^j
1	4.0	1	5.5	5.0	0.5	0.15	3.0
		2	4.0	6.0	0.7	0.12	3.5
2	4.0	1	5.5	15.0	0.5	0.15	3.0
		2	4.0	20.0	0.7	0.12	3.5
3	4.0	1	5.0	8.0	0.5	0.08	3.0
		2	8.0	10.0	0.7	0.03	3.5

The parameter settings considered for Study 1 is as follows. The first four settings are identical except for the reliability beliefs. The other settings include variations on other parameters as well.

Table 5.5: Suite of Parameter Settings in Study 1.

Setting No.	u^f	j	p_j	r_j	h_j	θ^j	c^j
1	4.0	1	5.5	5.0	0.5	0.80	3.0
		2	4.0	6.0	0.7	0.80	3.5
2	4.0	1	5.5	5.0	0.5	0.85	3.0
		2	4.0	6.0	0.7	0.90	3.5
3	4.0	1	5.5	5.0	0.5	0.90	3.0
		2	4.0	6.0	0.7	0.85	3.5
4	4.0	1	5.5	5.0	0.5	0.95	3.0
		2	4.0	6.0	0.7	0.95	3.5
5	4.0	1	5.5	15.0	0.5	0.85	3.0
		2	4.0	20.0	0.7	0.90	3.5
6	4.2	1	5.0	8.0	0.5	0.85	4.0
		2	8.0	10.0	0.7	0.90	4.0
7	4.5	1	7.0	8.0	0.9	0.95	3.8
		2	8.0	10.0	0.7	0.90	3.5
8	5.0	1	7.0	8.0	0.9	0.85	3.8
		2	8.0	10.0	0.7	0.85	3.5

5.10 Appendix C: Optimal Capacity Reservation Levels in Study 1

To find the optimal sourcing and contracting levels, let $\tilde{C}(q^1, q^2, q_1^f, q_2^f, \bar{Q^f})$ be the random variable denoting the one period cost of the firm if it reserves flexible backup capacity $\bar{Q^f}$ and orders q^j $(j = 1, 2)$ units from the dedicated supplier j and q_j^f $(j = 1, 2)$ units from the flexible backup supplier for product j. Then, if we let $i^j \in \{0, +\}$ denote the current state of the supplier j, where $i^j = 0$ if supplier j is down and $i^j = +$ otherwise, we have: $E_{D_2} E_{D_1}[\tilde{C}(q^1, q^2, q_1^f, q_2^f, \bar{Q^f})] =$

$$u^f \bar{Q^f} + \sum_{j=1}^{2} [c_j^f \, q_j^f + \mathbb{1}_{(i^j \neq 0)}(c^j q^j + G_j(q^j + q_j^f)) + \mathbb{1}_{(i^j = 0)} G_j(q_j^f)], \tag{5.91}$$

where $G_j(\cdot)$ is defined in (5.2). Hence, the expected cost as perceived by the firm is:

$$\begin{aligned} C_P(q^1, q^2, q_1^f, q_2^f, \bar{Q^f}) &= E_{i^2} E_{i^1} E_{D_2} E_{D_1}[\tilde{C}(q^1, q^2, q_1^f, q_2^f, \bar{Q^f})] \\ &= u^f \bar{Q^f} + \sum_{j=1}^{2} [c_j^f \, q_j^f + \theta^j (c^j q^j + G_j(q^j + q_j^f)) \\ &+ (1 - \theta^j) G_j(q_j^f)], \end{aligned} \tag{5.92}$$

where the subscript P on $C(\cdot)$ describes that it is the perceived (and not the true) value.

The problem for the firm then is to optimize the ordering and contracting decisions to minimize its *perceived cost* subject to the terms of the contract:

$$\min_{q^1, q^2, q_1^f, q_2^f, \bar{Q^f}} \quad C_P\,(q^1, q^2, q_1^f, q_2^f, \bar{Q^f}) \tag{5.93}$$

Subject to:

$$q_1^f + q_2^f \leq \bar{Q^f} \tag{5.94}$$

$$q^j, q_j^f \geq 0 \qquad\qquad (j = 1, 2). \tag{5.95}$$

Moreover, using Eq. (5.91), the *true* expected cost of the firm based on any given ordering and contracting decisions can be computed using the true reliabilities as:

$$C_T\left(q^1, q^2, q_1^f, q_2^f, \bar{Q}^f\right) = u^f \bar{Q}^f + \sum_{j=1}^{2} \left[c_j^f\, q_j^f + \left(1-\pi_0^j\right)\left(c^j q^j + G_j(q^j + q_j^f)\right) + \pi_0^j G_j(q_j^f)\right]. \tag{5.96}$$

Note that while the firm's decisions are based on its perceived reliabilities, the true cost defined in (5.96) depends both on the perceived and true reliabilities. In fact, we need to solve model (5.93)-(5.95) to derive the firm's optimal perceived decisions and implement them in Eq. (5.96) to determine the associated true total cost.

To solve our nonlinear model (5.93)-(5.95), we first need the following lemma.

Lemma 7 *The objective function* $C_P\left(q^1, q^2, q_1^f, q_2^f, \bar{Q}^f\right)$ *is jointly convex in its variables. Moreover, with* $\theta^j \in (0,1)$ $(j=1,2)$ *and* $u^f \neq 0$*, the convexity is strict.*

Proof of Lemma 7: To show convexity, we need to show that the Hessian matrix (H) of the function $C_P(\cdot)$ is positive semi-definite. Using Leibniz rule, $G_j'(x) = \left(h_j + p_j + r_j\right) F(x) - \left(r_j + p_j\right)$ and $G_j''(x) = \left(h_j + p_j + r_j\right) f(x)$. Therefore, the Hessian matrix can be written as

$$H = \begin{bmatrix} a_1 & 0 & a_1 & 0 & 0 \\ 0 & a_2 & 0 & a_2 & 0 \\ a_1 & 0 & b_1 & 0 & 0 \\ 0 & a_2 & 0 & b_2 & 0 \\ 0 & 0 & 0 & 0 & u^f \end{bmatrix},$$

where $a_j = \theta^j \left(h_j + p_j + r_j\right) f_j(q^j + q_j^f)$ and $b_j = a^j + \left(1-\theta^j\right) f(q_j^f)$ $(j=1,2)$. Next, solving the characteristic equation $|H - \lambda I| = 0$, we obtain the following eigenvalues:

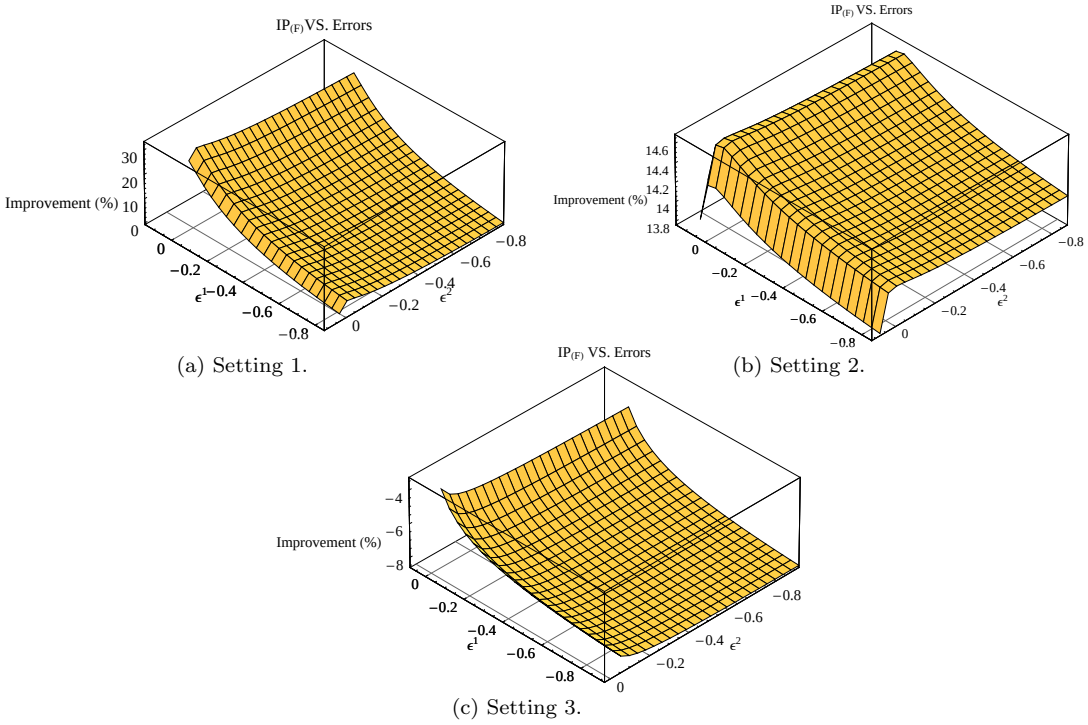
$$
\begin{aligned}
\lambda_1 &= \frac{1}{2}(h_1+p_1+r_1)\left[(1-\theta^1)f_1(q_1^f)+2\,\theta^1 f_1(q^1+q_1^f)\right.\\
&\qquad\left.+\sqrt{[(1-\theta^1)f_1(q_1^f)]^2+[2\,\theta^1 f_1(q^1+q_1^f)]^2}\right],\\
\lambda_2 &= \frac{1}{2}(h_1+p_1+r_1)\left[(1-\theta^1)f_1(q_1^f)+2\,\theta^1 f_1(q^1+q_1^f)\right.\\
&\qquad\left.-\sqrt{[(1-\theta^1)f_1(q_1^f)]^2+[2\,\theta^1 f_1(q^1+q_1^f)]^2}\right],\\
\lambda_3 &= \frac{1}{2}(h_2+p_2+r_2)\left[(1-\theta^2)f_2(q_2^f)+2\,\theta^2 f_2(q^2+q_2^f)\right.\\
&\qquad\left.+\sqrt{[(1-\theta^2)f_2(q_2^f)]^2+[2\,\theta^2 f_2(q^2+q_2^f)]^2}\right],\\
\lambda_4 &= \frac{1}{2}(h_2+p_2+r_2)\left[(1-\theta^2)f_2(q_2^f)+2\,\theta^2 f_2(q^2+q_2^f)\right.\\
&\qquad\left.-\sqrt{[(1-\theta^2)f_2(q_2^f)]^2+[2\,\theta^2 f_2(q^2+q_2^f)]^2}\right],\\
\lambda_5 &= u^f.
\end{aligned}
$$

Now notice that since (1) $f_j(\cdot)$ is a probability density function, (2) $\theta^j \in [0,1]$, and (3) $r_j, p_j, h_j, u^f \in [0,+\infty)$, all the above eigenvalues are nonnegative. Therefore H is positive semi-definite and hence $C_P(\cdot)$ is jointly convex. Moreover, if $\theta^j \neq 0,1$ and $u^f \neq 0$, all the eigenvalues are positive; hence, H is positive definite. Thus, with $\theta^j \neq 0,1$ and $u^f \neq 0$, $C_P(\cdot)$ is also *strictly* convex. □

Study 4 Consider a firm facing two Normally (and independently) distributed demands for its products. Particularly, let D_1 and D_2 respectively follow $N(5000, 1200^2)$ and $N(3000, 800^2)$, where $N(\mu, \sigma^2)$ denotes a Normal distribution with mean μ and a standard deviation of σ. To illustrate different cases, we consider three sets of different parameter settings presented in Table 5.4 in Appendix B. Fig. 2 depicts the corresponding Improvement Percentages (IP's) in the firm's true expected costs due to contracting with the flexible supplier versus its reliability belief error $\Upsilon = (\epsilon^1, \epsilon^2)$.

We denote by $IP_{(F)}\%$ the cost improvement percentage due to the existence of the flexible backup supplier: $IP_{(F)}\% = \frac{V^f}{|C_T\,(q'^{1*},q'^{2*},0,0,0)|} \times 100$. Note that $IP_{(F)}\% > 0$ implies that contracting is profitable and $IP_{(F)}\% \leq 0$ indicates a non-profitable contracting situation. The former case can be seen in parts (a) and (b), and the latter can be seen in part (c) of Fig. 2. The corresponding capacity reservation levels for each parameter setting are depicted in Fig. 1.C (in Appendix C). As one specific example, a firm with parameter setting 1 (see Table 5.4 in Appendix B) and with a reliability belief of $\Theta = (0.8, 0.9)$ (i.e., with belief error $\Upsilon = (-0.05, +0.02)$ based on π^1 and π^2 presented in Table 5.4) will form a contract and reserve a capacity level of $\bar{Q^f}^* = 6238.9$ units. Based on this decision, the firm would be able to reduce its expected total true costs by $IP_{(F)}\% = 29.5\%$. However, as can be seen in Fig. 2 part (a), if this firm has large errors in its reliability belief, it will not be able to greatly reduce its costs by contracting with the flexible backup supplier. (For instance, $IP_{(F)}$ is less than 1% with $\Upsilon \approx (-0.8, -0.8)$.) However, a firm with parameter setting 2 and with the same reliability belief ($\Theta = (0.8, 0.9)$) will reserve a capacity of $\bar{Q^f}^* = 8945.5$ units and will be able to reduce its expected total true costs by $IP_{(F)}\% = 14.6\%$. Although the percentage benefit for this firm is less than the first one, even with large errors in its reliability belief (e.g. with $\Upsilon \approx (-0.8, -0.8)$), as can be seen in Fig. 2 part (b), it will still be able to reduce its true expected costs approximately by 14%. In other words, accuracy in estimating the dedicated suppliers' reliabilities is critical for the former firm, but not for the latter (see also Study 4 for more details). This is due to the high profit margins in setting 2 which makes a secondary backup flexible supplier still highly valuable, even with large errors in belief. This results in Observation 10 presented in the main body of the chapter.

Study 5 Consider a firm facing two Normally distributed demands for its products as discussed in Study 4. Let $IP_{(I)}\%$ denote the Improvement Percentage (IP) in the firm's expected true costs due to obtaining information about disruption risk of both of its unreliable suppliers. This value can be computed by $IP_{(I)}\% =$



(a) Setting 1.

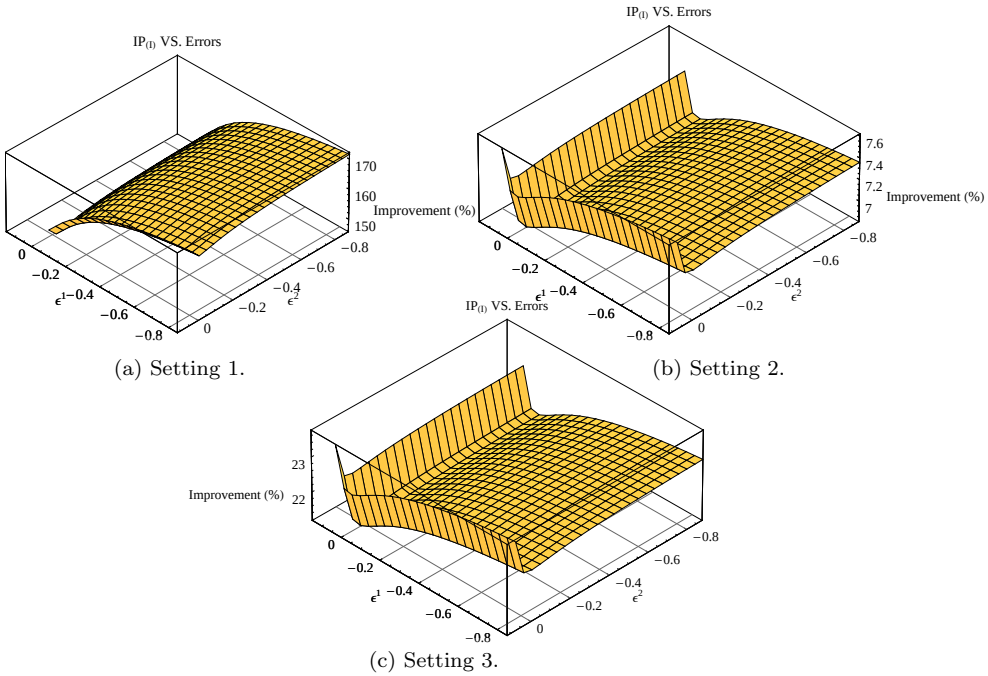
(b) Setting 2.

(c) Setting 3.

Figure 5.2: The Value of a Flexible Backup Supplier ($IP_{(F)}\%$) in Settings 1-3 for Different Values of Error in the Firm's Reliability Belief (ϵ^1, ϵ^2)

.

$\frac{VI}{|C_T\,(q^{1*},q^{2*},q_1^{f*},q_2^{f*},\bar{Q}^{f*})|} \times 100$, where $C_T\,(q^{1*}, q^{2*}, q_1^{f*}, q_2^{f*}, \bar{Q}^{f*})$ is the firm's expected true cost under its perceived optimal decisions before obtaining information. Fig. 3 illustrates different values of $IP_{(I)}\%$ for the parameter settings 1-3 (presented in Table 5.4 in Appendix B) versus the errors in the firm's reliability belief. As some particular examples, considering the cases discussed in Study 4. A firm with parameter setting 1 and with a reliability belief vector of $\Theta = (0.8, 0.9)$ ($\Upsilon = (-0.05, +0.02)$) can greatly reduce its true costs by $IP_{(I)}\% = 148.9\%$, if it obtains full-information about its unreliable suppliers' disruption risk. However, a firm with parameter setting 2 and with same belief (and same error as the previous firm) $\Theta = (0.8, 0.9)$ ($\Upsilon = (-0.05, +0.02)$) can only reduce its cost by $IP_{(I)}\% = 6.9\%$. Finally, a firm with parameter setting 3 and with $\Theta = (0.87, 0.90)$ ($\Upsilon = (-0.05, -0.07)$) can benefit from full-information



(a) Setting 1.

(b) Setting 2.

(c) Setting 3.

Figure 5.3: The Value of Information ($IP_{(I)}\%$) for Different Values of Firm's Reliability Belief Error (ϵ^1, ϵ^2).

by an amount equal to $IP_{(I)}\% = 19.2\%$ of its current (no-information) true cost. Therefore, if this latter firm is risk-neutral and if establishing a system to obtain such information adds its costs by 10%, it should choose to establish such a system and hence reduce its true expected total cost by 9.2%.

This study reveals additional interesting insights. First, notice that the information is much more valuable in setting 1 than setting 2. Since setting 1 represents lower profit margins than setting 2, this suggests Observation 11 presented in the main-body of the chapter.

This observation is consistent with what we observed in Study 4 and can be explained as follows. When profit margins are tight, overinvesting in the expensive secondary supplier (resulting from underestimating the reliabilities) is very costly, and cannot be justified by the profit obtained from higher potential sales in the case of a disruption. On the other hand, underinvesting in the flexible resource

(resulting from overestimating the reliabilities) is more harmful in the case of high profit margins than low ones only when a disruption actually occurs and because of the lost sales. However, the probability of facing a disruption is low. Thus, in expectation, underinvesting is also percentage-wise more costly when profit margins are low than when they are high.

Second, as can be seen in Fig. 3, when a firm is overestimating the reliability of its primary supplier (i.e., when $\epsilon^j > 0$), the value of information is much more sensitive to belief errors than the case of underestimating. This results in Observation 12 presented in the main body of the chapter.

Notice that when a firm overestimates the reliability of its suppliers, it invests less in the secondary flexible backup capacity. In such a situation, with a little better estimation (less error), the firm will invest a little more in the backup capacity. Although the change in investment is relatively small, the cost benefit is large since the backup capacity can work as an *insurance* in the event of pernicious disruptions.

5.11 Appendix D: Further Results on the Two-Product Setting with Recourse

In this appendix, we derive the optimal ordering behavior of the firm (under the two-product setting with recourse) in Stage 2 for the case where $c_j^f > c^j$ for both $j = 1, 2$. Note that the analysis in this case is much simpler than the case $c_j^f \leq c^j$ provided in the main body of this chapter.

We start with the case where the firm observes both unreliable suppliers to be up.

Proposition 14 (Both Suppliers up) *Given any reserved flexible backup capacity of $\bar{Q}^f$, when $c_j^f > c^j$, it is optimal to set $q_j^{f*} = 0$ and $q^{j*} = F^{-1}(\frac{p_j + r_j - c^j}{p_j + r_j + h_j})$ when both unreliable suppliers are observed to be up.*

When the firm only observes one of the unreliable suppliers to be up, the optimal

ordering policy is as follows.

Proposition 15 (One Supplier up) *Let $m \in \{1,2\}$ denote the dedicated supplier that is observed to be up, and $n = 3 - m$ be the disrupted supplier. Given a reserved flexible backup capacity of $\bar{Q}^f$, when $c_j^f > c^j$, it is optimal to set $q_n^{f*} = \min\{\bar{Q}^f, F^{-1}(\frac{p_j+r_j-c_n^f}{p_j+r_j+h_j})\}$, $q_m^{f*} = 0$, and $q^{m*} = F^{-1}(\frac{p_j+r_j-c^m}{p_j+r_j+h_j})$.*

When the firm observes both unreliable suppliers to be disrupted, the ordering policy is the same as Theorem 9. From these results we have the following:

Observation 18 Unlike the case studied in the main body, when $c_j^f \geq c^j$, rationing the limited backup capacity only occurs when both unreliable suppliers are observed to be down.

Proof of Proposition 14: Please note that, as before, the KKT conditions are necessary and sufficient to characterize the optimal solution. Moreover, the KKT conditions are the same as those in the proof of Theorem 7:

$$q_1^f + q_2^f \leq \bar{Q}^f \tag{5.97}$$

$$c^j - \lambda^j = \mu - \lambda_j^f + c_j^f \qquad (j = 1,2) \tag{5.98}$$

$$q^j + q_j^f = F_j^{-1}(\frac{p_j + r_j - (c^j - \lambda^j)}{p_j + r_j + h_j}) \qquad (j = 1,2) \tag{5.99}$$

$$\mu(q_1^f + q_2^f - \bar{Q}^f) = 0 \tag{5.100}$$

$$\lambda^j q^j = 0 \qquad (j = 1,2) \tag{5.101}$$

$$\lambda_j^f q_j^f = 0 \qquad (j = 1,2) \tag{5.102}$$

$$q^j, q_j^f, \mu, \lambda^j, \lambda_j^f \geq 0. \qquad (j = 1,2) \tag{5.103}$$

Observe that setting $\lambda^j = 0$, $\mu = 0$, $\lambda_j^f = c_j^f - c^j$, $q_j^{f*} = 0$, and $q^{j*} = F^{-1}(\frac{p_j+r_j-c^j}{p_j+r_j+h_j})$ satisfy the above KKT conditions. □

Proof of Proposition 15: Note that the KKT conditions for this case are the same as those in the proof of Theorem 8:

$$q_1^f + q_2^f \leq \bar{Q^f} \tag{5.104}$$

$$c^m - \lambda^m = \mu - \lambda_m^f + c_m^f \tag{5.105}$$

$$q^m + q_m^f = F_m^{-1}(\frac{p_m + r_m - (\mu - \lambda_m^f + c_m^f)}{p_m + r_m + h_m}) \tag{5.106}$$

$$q_n^f = F_n^{-1}(\frac{p_n + r_n - (\mu - \lambda_n^f + c_n^f)}{p_n + r_n + h_n}) \tag{5.107}$$

$$\lambda^m q^m = 0 \tag{5.108}$$

$$\lambda_m^f q_m^f = 0 \tag{5.109}$$

$$\lambda_n^f q_n^f = 0 \tag{5.110}$$

$$q^m, q_m^f, q_n^f, \mu, \lambda^m, \lambda_m^f, \lambda_n^f \geq 0. \tag{5.111}$$

If $\bar{Q^f} \geq F^{-1}(\frac{p_j + r_j - c_n^f}{p_j + r_j + h_j})$, set $\mu = 0$. Otherwise, chose μ such that $F^{-1}(\frac{p_j + r_j - c_n^f - \mu}{p_j + r_j + h_j}) = \bar{Q^f}$. Next, observe that setting $\lambda_m^f = \mu + c_m^f - c^m$, $\lambda_n^f = 0$, $\lambda^m = 0$, $q_n^{f*} = \min\{\bar{Q^f}, F^{-1}(\frac{p_j + r_j - c_n^f}{p_j + r_j + h_j})\}$, $q_m^{f*} = 0$, and $q^{m*} = F^{-1}(\frac{p_j + r_j - c^m}{p_j + r_j + h_j})$ satisfy the KKT conditions, which are sufficient and necessary to characterize the optimal solution. □

CHAPTER 6

Supply Chain Disruption Risk Management: Dynamic Analyses

6.1 Introduction

We have observed many pernicious supply disruptions occurring in different industries in recent years. Perhaps the most recent one is the 2011 earthquake and tsunami in Japan that affected virtually all of the automakers in the world. Learning from such events, companies try to adopt better strategies for disruption risk mitigation. One obvious strategy is to have backup suppliers for different products. For instance, according to Reuters, after the recent 2011 disaster in Japan, Toyota is taking several steps to reduce its exposure to such disasters, and the first one is to make sure that the items have backup suppliers ([89]). This can be achieved for instance by standardizing parts across suppliers and users or by establishing backup supplier/manufacuturing capacities. Another recent example of establishing backup capacities can be seen in the case of disruption in the German company, Merck, that produces 100% of supply of a pearl-luster pigment called Xirallic ([37]). According to Wall Street Journal, the production plant located at the northeastern Japan was hit hard by the March 2011 quake and caused problems for many of the auto makers ([37]). Learning from the disaster, the company is now setting up a second production

line in Germany ([37]).

A second rather obvious mechanism to implement is to carry more inventory so that disruptions do not raise havoc in the supply chain. As an example, a second step taken by Toyota after the recent 2011 earthquake in Japan, according to Reuters, is to ask suppliers to hold extra inventory ([89]). Another example of the use of this strategy can be seen in the case of Johnson & Johnson that provides medical supplies for the Pentagon. Since the Pentagon knows that in the case of a war or a major disaster it will require huge amounts of medical supplies, Johnson & Johnson is under contract by the U.S. government to maintain certain inventory levels of medical supplies ([129]). Carrying inventory, however, can be very costly for firms, and is commonly thought to violate the principles of lean production.

A third mechanism that can substitute stockpiling inventory or establishing separate backup capacities for each of the products is to inject flexibility into the backup system (see, e.g., [122]). The benefit of flexibility in the backup system can be observed from the case of the disruption in the Toyota supply chain in 1997 when a fire at the Aisin Seiki Co. destroyed most of the capacity to manufacture P-valves. According to the Wall Street Journal, Toyota officials called different suppliers to obtain P-valves, including Somic ([115]). Somic did not have any contract to allocate part of its production capacity to produce P-valves, but it did have the flexibility to do so. It flexed its production line to make P-valves and it delivered its first P-valves to Toyota right on time ([115]).

A forth mechanism that is used in practice is to monitor suppliers in an effort to create visibility by obtaining some information about the progression of disruption risk over time. An example of using this mechanism is the practice of the Critical Material Council (CMC) that monitors the semiconductor material supply chain. It watches the shifting patterns in capital expenditures and R&D investments and look for potential shortages in supply ([129]). It provides a venue of information sharing between material suppliers and material customers, and develops alternative sources

of supply when there is a potential for a disruption.

Disruption risk information can be obtained in several different ways. For instance, Open Rating is a leading provider of supply risk management solutions that enables firms to collect disruption risk information and adopt suitable actions. The benefit of monitoring suppliers and obtaining disruption risk information can be understood from the disruption case in Boeing's inflexible supply chain in 2007. Advanced Integration Technology (AIT) faced a disruption and fell behind supplying parts needed to assemble Boeing's 787 (Dreamliner). Evidence showed that AIT had been facing serious production problems in the year before (see, for instance, [55]), and if Boeing had carefully monitored AIT, it might have made better procurement decisions ([122]).

In this chapter, we consider a firm that procures from several suppliers, and study the above mentioned practices in mitigating disruption risks. We provide insights into the effectiveness of such mechanisms as well as their interplay. For realism, we allow for disruption risks to dynamically change over time. We model dynamics of disruption risks as a Discrete Time Markov Chain (DTMC) with several threat levels indicating the "health level" of suppliers. As an example, the S&P credit risk rating system with states {1=AAA, 2=AA, 3=A, 4=BBB, 5=B/BB, 6=CCC/CC/C} $\cup$ $\{0 = Default\}$ is an analogous system for which Markov chain modeling is commonly used. These threat levels may also represent weather forecasts, potential economical sanctions, security levels, etc. It is noteworthy that our Markov model is a step forward from the prevalent assumption of i.i.d. Bernoulli disruptions in the literature which does not model the effect of the length of a disruption.

We consider a firm that procures several products from unreliable suppliers, and first analyze the case in which the firm can perfectly monitor its suppliers, i.e., a case of full-information on disruption threat levels. We show how to quantify and optimize several interrelated policy issues: an upfront investment level in a flexible backup capacity, dynamic ordering from different suppliers considering their disruption threat

levels, and carrying inventory of different products (over time). We next consider the case in which the firm cannot effectively monitor its unreliable suppliers, and hence, does not have information on their disruption risk levels. By differentiating between the firm's perception and reality, we generate several insights for firms that suffer from lack of disruption risk information (disruption visibility) in their supply system. Furthermore, comparing scenarios with and without disruption risk information, we quantify the value of monitoring suppliers. We also consider the value of flexibility in the backup system in the sense of using a single flexible backup supplier instead of multiple dedicated backup ones. Our study results in several insights for both researchers and practitioners that we summarize in Section 6.5.

The remainder of the chapter is organized as follows. We review the literature in the next section, then in Section 6.3 we describe our model. Section 6.4 presents our analyses under two cases of information availability, and provides several insights into the value of different disruption risk mitigation mechanisms. Finally, Section 6.5 summarizes the insights gained and concludes.

6.2 The Literature

The awareness of the damage done by disruptions has motivated an increasing number of papers that study different mechanisms to mitigate disruption risks. Disruption risks can be modeled as either static (e.g. single-shot or repeated settings) or dynamic events, and can be classified in eight categories: dynamic or static random disruptions (i.e., all-or-nothing), dynamic or static random yield, dynamic or static random capacity, and dynamic or static financial default. In this chapter, we consider the first category, i.e., dynamic random disruptions.

For studies that consider the case of random disruptions, we refer interested readers to [113], [58], [108] and [38] [145], [13], [122], and the references therein. Studies that consider the case of random yield include [48], [10], [1], [140], [59], [43], [32],

and [23], and a review can be found in [41]. For the stream of research that deals with random capacity in the supply system, we refer readers to studies such as [27], [24], [20], and the references therein. Examples of modeling disruptions as financial defaults can be found in [14], [13], and [141]. Finally, some studies including [158] consider the effect of allowing for a few of the above-mentioned types of disruptions.

While in reality most disruption risks evolve dynamically over time, most of the above-mentioned studies have focused on the static disruption cases. However, there are a few studies in the literature that consider dynamics of disruptions. Authors of [146], for instance, develop multi-period models with dynamic disruptions in which the firm has a single unreliable supplier, as well as models in which a second, perfectly reliable supplier is available. In addition to such studies, the inventory control literature with Markovian supply availability is also relevant to our study, although it usually studies single-sourcing settings without any supply flexibility. Within this literature, [133] presents a fundamental and excellent study with periodic review inventory control where information about the evolution of the supply system is modeled as a Markov chain. [112] addresses a periodic-review setting with setup costs, where the probability that an order placed now is filled in full depends on whether supply was available in the previous period . We contribute a new perspective by modeling both complete and incomplete information about the dynamic disruption process, and by allowing for flexibility in the backup system. We compare scenarios with and without disruption state information to obtain the value of monitoring suppliers. We also capture the value of flexibility, and its interactions with information.

Some part of the literature models disruptions as static events, including multi-period models but with repeated disruptions. [144] uses a Bayesian approach for supply learning (i.e., reliability-forecast updating) with i.i.d. Bernoulli disruptions and characterizes the firm's optimal sourcing and inventory decisions. [10] studies a finite-horizon, discrete-time, i.i.d. stochastic continuous demand model in which there are two zero lead time random-yield suppliers.

The literature on supply-disruption research can also be viewed from the perspective of single versus multi-supplier models. Single supplier studies include [104], [19], [111], [112], [57], [133], [107], [110], [108], and [38]. Multi supplier studies can be found in [10], [113],[58], [14], [13], [144], [146], [32], [158], and [122]. In this chapter, we allow for an arbitrary number of products/suppliers to generate insights for firms that procure several items from various unreliable supplier.

In addition to considering a multi-period dynamic disruption model with arbitrary number of suppliers, another distinct feature of our modeling framework is that we allow for product mix flexibility in the backup system, and study how such flexibility can be used to mitigate disruptions. Reviews of flexibility can be found in studies such as [128], [49], and [139]. Operational mix flexibility has been studied in papers such as [46], [85], [154], [93], [53], [147], [82], and [122]. Our study contributes to this literature by considering the value of a flexible backup supplier/resource to compensate for unreliability of dedicated suppliers, particularly in the more complex dynamic setting. Roughly similar contributions in the context of control of flexible queueing networks can be found in [6] and [124].

6.3 The Model

Consider a firm that produces/sells $n = |N|$ types of products (where $N = \{1, 2, \cdots, n\}$ denotes the set of underlying products) and has n dedicated unreliable suppliers, each capable of supplying an important component (or raw material) for one of the products. We assume the production and inventory levels are continuous. Denote the dedicated supplier of raw material of product $j \in N$ as supplier j. To insure the supply stream against future disruptions, the firm can also establish (or contract with) a flexible backup resource, namely f, at a limited capacity $\bar{Q}^f \in (0, \infty)$ that can produce quantities of underlying products, the sum of which cannot exceed $\bar{Q}^f$. Establishing such a capacity, however, is costly. Let $g(u^f, \bar{Q}^f)$ denote the invest-

ment cost at the flexible backup capacity, which depends on the capacity level $\bar{Q}^f$ as well as the "per unit" investment cost u^f. We allow for a general class of investment costs represented through the cost function $g(u^f, \bar{Q}^f)$. However, to represent a "well-behaved" investment cost function, we assume $g : \mathbb{R}^2_+ \to \mathbb{R}_+$ is continuous, increasing in u^f with $g(0, \cdot) = 0$, increasing convex in capacity $\bar{Q}^f$ with $g(\cdot, 0) = 0$, and supermodular (twice differentiable with positive cross partials). We note that a special case of this type of investment is that of reserving some backup capacity through a *capacity reservation* contract (also known as an option contract), where paying an up-front cost of $g(u^f, \bar{Q}^f) = u^f \bar{Q}^f$ will let the firm to reserve a backup capacity of $\bar{Q}^f$ units. This type of contract is prevalent in many industries including semiconductors, consumer electronics, telecommunications, and pharmaceutical (where the demand is highly volatile and difficult to forecast), automakers, textile and garment industry (for more details on this contract see [127], [122], and the references therein). For convenience, we use subscripts for products, superscripts for suppliers, and employ the following notation, where each variable is in $[0, \infty)$:

h_j : Holding cost per unit of product j per period; $(j \in N)$

p_j : backorder penalty cost per unit of unmet demand of product j; $(j \in N)$

c^j : Per unit purchasing cost of product j from dedicated supplier j; $(j \in N)$

c^f_j : Per unit purchasing cost of product j from the flexible backup supplier; $(j \in N)$

u^f : Per unit capacity reservation cost of the flexible backup supplier;

q^j : Order quantity from dedicated/primary supplier j; $(j \in N)$

q^f_j : Order quantity from flexible backup supplier for product j; $(j \in N)$

$\bar{Q}^f$: Reserved capacity from flexible backup supplier.

Fig. 6.3 depicts the two-echelon supply chain model under consideration. The firm operates a periodic-review inventory system where unmet demand is backlogged, and it has to pay the purchasing cost c^j and c^f_j per order of product $j \in N$ delivered by dedicated (and unreliable) supplier j and the flexible backup resource, respectively. The flexible backup resource has a shared and limited capacity; it can deliver any

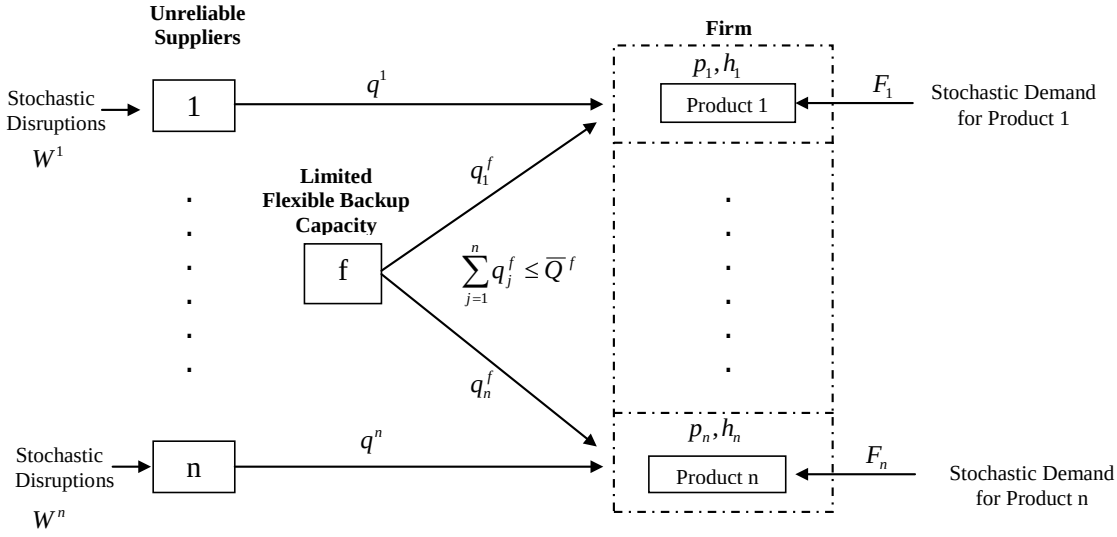


Figure 6.1: The general model under consideration.

combination of quantity of products $(q_j^f : j \in N)$ as long as $\sum_{j \in N} q_j^f \leq \bar{Q}^f$, when it is established (or contracted) at a capacity level of $\bar{Q}^f$. To match reality and eliminate trivial cases, we shall assume none of the products in set N can be procured for free; $c_j^f + u^f > 0$ and $c^j > 0$ for $j \in N$.

For $j \in N$, let $\mathcal{L}_j(x) = h_j[x]^+ + p_j[-x]^+$ and define the expected one-stage cost

$$G_j(x) = E_{D_j}[\mathcal{L}_j(x - D_j)] = h_j \int_{-\infty}^{x} (x - \xi) dF_j(\xi) + p_j \int_{x}^{\infty} (\xi - x) dF_j(\xi), \qquad (6.1)$$

where $[x]^+ = \max\{0, x\}$, and $F_j(\cdot)$ is the cumulative distribution function (c.d.f) of the demand for product j, D_j. We assume that demands for each product j across periods are i.i.d. random variables, and further, D_j and $D_{j'}$ are independent (for all $j, j' \in N$ s.t. $j \neq j'$). For the most part of the chapter, we consider the case where unmet demand in a period is backlogged. However, we will also investigate a closely related model in the numerical analyses (at the end of Section 6.4) where we assume unmet demand is lost. The above model with backordering and penalties for such is very similar conceptually to one with lost sales and a revenue per unit sold. While they are different in the precise details of the inventory policies, for the purposes of

our study, they are not much different with respect to the themes of the value of information and backup supplier flexibility. The backordering formulation is superior for analysis. On the other hand, the smaller state space of the lost sales model makes it more tractable for numerical experiments so we use it for that purpose. However, it should be noted that the computational methodology we develop does not differ significantly between the two.

We model the disruption risk processes of the dedicated suppliers via a discrete time Markov process. Let s^j denote the threat level of dedicated supplier j (as an indicator of its health), where $s^j = 0$ means dedicated supplier j is in the down (default) state and $s^j = k > 0$ denotes that it is in threat level k. We assume that the dynamics of disruptions can be modeled as a Discrete Time Markov Chain (DTMC) with state space $S^j = \{0, 1, \ldots, k^j\}$ for dedicated supplier j. Let $\mathbf{W}^j = [w^j_{lm}]$ denote the transition probability matrix of DTMC of supplier j, where w^j_{lm} is the probability that it will be in risk level m in the next period given that the current risk level is l. The set $W = \{\mathbf{W}^j, j \in N\}$ completely describes the dynamics of disruptions of unreliable suppliers. For every $j \in \mathcal{N}$, we assume $w^j_{k0} < w^j_{k'0}$ for every $0 < k < k'$ in S^j (i.e., the higher the threat level, the higher the risk of disruption). We assume every element of W is *aperiodic* and *irreducible*. Thus the underlying DTMC's are all *ergodic* and have a steady state distribution which for supplier j we denote by the vector $\pi^j = (\pi^j_0, \pi^j_1, \ldots, \pi^j_{k^j})$. Hence, π^j_0 is the long-run disruption probability of dedicated supplier j and $(1 - \pi^j_0)$ is its reliability: the long-run fraction of time that it is not disrupted. When the information on the threat levels are not perfect, we denote the firm's (steady-state) reliability perception error about unreliable supplier j by ϵ^j, and let $\Upsilon = (\epsilon^j, j \in N)$ denote the vector of reliability perception errors.

6.4 Analyses

As discussed in the Introduction, one way to mitigate the risk of supply disruptions is to carry inventory over time. This practice is, however, costly. Hence, there is a trade-off for a firm between investing in a backup supplier and carrying inventory. Also, disruption risk of a supplier may change over time. Moreover, firms in practice pay attention to the length of a disruption (in addition to its probability of occurring). These latter observations highlight the need to carefully develop a dynamic setting. Such observations cannot be captured even if one models disruptions as i.i.d. Bernoulli variables (which is prevalent in literature).

We contribute a modeling and analysis framework in which disruption risk levels may change dynamically over time. We assume that the firm has an option to establish (or reserve) a backup flexible capacity at time 0 to insure the supply stream against future disruptions. The firm then exercises a periodic review inventory control in every future period during which it can procure either from the primary suppliers or from the reserved secondary flexible capacity. Unmet demand is backordered and supply lead times and production cycles are negligible in comparison with the review period. We assume the following order of events within each review period: (1) The firm observes the state of the system subject to an information constraint. (2) The firm decides the order sizes and orders from all suppliers subject to the contracts. (3) Product demands are realized. (4) Holding cost or shortage cost accrue. (5) The state of the system is updated, including the inventory and disruption risk levels. In this setting, we capture the value of the flexible supplier and consider different scenarios of information availability, where the disruption risk levels of suppliers may or may not be observable by the firm. Comparing these scenarios reveals the value of disruption risk information for the firm.

Let the vector $\mathbf{x}(t) = (x_j(t) : j \in N)$ denote the inventory on hand of the underlying products at period t. Also, let $\mathbf{q}(t) = (q^j(t) : j \in N)$ and $\mathbf{q}^f(t) = (q^f_j(t) : j \in N)$, respectively, denote the vectors of order sizes from the primary suppliers and

the flexible supplier at period t. Additionally, let $\beta \in (0,1)$ be the discount factor and $\mathbf{s}(t) = (s^j(t) : j \in N, s^j(t) \in S^j)$ denote the state of disruption risk levels of the dedicated suppliers at period t. In practice, risk levels of unreliable suppliers are less than perfectly assessable by firms (i.e., the vector $\mathbf{s}(t)$ may not be observable for a firm). Obtaining such information requires a careful monitoring of suppliers. Key questions include the value of information, the value of a flexible backup supplier, and the interplay between flexibility and information. To answer such questions, we first analyze and discuss the full-information scenario where the vector of disruption threat levels, $\mathbf{s}(t)$ (for $t = 1, 2, ...$) is observable (and assessable) for the firm. Then we consider the case where this vector is not observable and develop a Partially Observable Markov Decision Process (POMDP). Comparing these scenarios reveals the benefit of monitoring unreliable suppliers.

6.4.1 Full Information on Disruption Risk Levels

Let $\tilde{J}(\mathbf{x}(0), \mathbf{s}(0))$ denote the optimal expected infinite-horizon discounted cost of the firm (including the investment cost at period $t = 0$) if the initial disruption risk levels are $\mathbf{s}(0)$ and the firm starts with an inventory on-hand vector of $\mathbf{x}(0)$. This value can be computed by the following program:

$$\tilde{J}(\mathbf{x}(0), \mathbf{s}(0)) = \min_{\bar{Q}^f \in \mathbb{R}_+} g(u^f, \bar{Q}^f) + J_{\bar{Q}^f}(\mathbf{x}(0), \mathbf{s}(0)), \tag{6.2}$$

where $J_{\bar{Q}^f} : \mathbb{R}^n \times (\Pi_{j \in N} S^j) \to \mathbb{R}_+$ is the optimal infinite-horizon discounted cost of the firm given the established capacity $\bar{Q}^f$. $J_{\bar{Q}^f}(\cdot, \cdot)$ can be computed using the

following Bellman equation for all $t \in \mathbb{Z}^+$: $J_{\bar{Q}^f}(\mathbf{x}(t), \mathbf{s}(t)) =$

$$\min_{\mathbf{q}(t), \mathbf{q}^f(t) \geq 0 : \sum_{j \in N} q_j^f(t) \leq \bar{Q}^f} \Big\{ \sum_{j \in N} \big[c_j^f q_j^f + \mathbb{1}_{(s^j(t)>0)} (c^j q^j(t) + G_j(x_j(t) + q^j(t) + q_j^f(t))) + \mathbb{1}_{(s^j(t)=0)} G_j(x_j(t) + q_j^f(t)) \big] + \beta \, \mathbb{E}_{\mathbf{D}(t)} \mathbb{E}_{\mathbf{s}(t+1)} \left[J_{\bar{Q}^f}(\mathbf{x}(t+1), \mathbf{s}(t+1)) | \, \mathbf{x}(t), \mathbf{s}(t) \right] \Big\}, \tag{6.3}$$

where the inventory transition rule from $\mathbf{x}(t)$ to $\mathbf{x}(t+1)$ is:

$$\mathbf{x}(t+1) = \mathbf{x}(t) + \mathbf{q}^f(t) + (\mathbb{1}_{\{s^1(t) \geq 0\}}, \mathbb{1}_{\{s^2(t) \geq 0\}}, \cdots, \mathbb{1}_{\{s^n(t) \geq 0\}}) . \mathbf{q}(t) - \mathbf{d}(t), \tag{6.4}$$

and threat level transitions from $\mathbf{s}(t)$ to $\mathbf{s}(t+1)$ are defined through Markov processes governed by set of t.p.m's W.

Now, solving program (6.2)-(6.3) derives the firm's optimal expected infinite-horizon discounted cost as well as its optimal investment level under full-information. This in turn yields a measure for the true value of the flexible backup resource under full-information:

$$\Delta^f(\mathbf{x}(0), \mathbf{s}(0)) = J_0(\mathbf{x}(0), \mathbf{s}(0)) - \tilde{J}(\mathbf{x}(0), \mathbf{s}(0)).$$

For instance, $\Delta^f(\mathbf{x}(0) = \mathbf{0}_{1 \times n}, \mathbf{s}(0) = \mathbf{1}_{1 \times n})$ provides a good measure for investigating the value of the backup resource by setting all initial inventory levels to zero and placing all suppliers in their most reliable state.

We start our investigations by answering the following questions.

Question 9 How should firms update their inventory level safeguards to effectively mitigate against dynamic risks of disruption?

We notice that even without the unreliable suppliers and partial information, the problem of managing inventories with the existence of a shared limited capacity is

known to be difficult (see, e.g., [38]). The answer to the above question is hard in general; however, we are able to show that under some conditions a simple *state-dependent base-stock policy* is optimal. That is, the firm should set inventory safeguards for each of the products, where the inventory safeguard of product j depends on the state of the unreliable supplier of this product. Furthermore, under some conditions, for each product (a) the safeguard is independent of the threat levels of suppliers of other products, (b) in each period, the firm should reach the inventory safeguard either through the primary unreliable supplier or through the backup capacity, but not both (a single-sourcing strategy at any given period), and (c) the firm should implement higher safeguards when the unreliable supplier is in a higher threat level, assuming the process governing disruptions is monotone as formalized below.

To define a monotone disruption process, we first need the following definition.

Definition 4 (Stochastically Monotone DTMC, Daley [36]) *A DTMC* $\{X_n, n = 0, 1, \cdots\}$ *is defined to be* stochastically monotone *when its one-step transition probability function* $Pr\{X_{n+1} \leq y | X_n = x\}$ *is non-increasing in* x *for every fixed* y*.*

We next consider a relabeling of states so that the disruption state, state 0, of a k-level threat system is considered to be state k+1. Specifically, for the DTMC $\{s_n^j, n = 0, 1, \cdots\}$ defined on space $S^j = \{0, 1, \ldots, k^j\}$, consider the relabeled DTMC $\{\tilde{s}_n^j, n = 0, 1, \cdots\}$ defined on space $\tilde{S}^j = \{1, \ldots, k^j, k^j + 1\}$, where $\tilde{s}_n^j = s^j \mathbb{1}_{(s^j > 0)} + (k^j + 1)\mathbb{1}_{(s^j = 0)}$. We denote the relabeled version of the t.p.m. $\mathbf{W}^j = [w_{lm}^j]$ by $\tilde{\mathbf{W}}^j = [\tilde{w}_{lm}^j]$. Finally, using the above definition, we say that the relabeled DTMC governed by $\tilde{\mathbf{W}}^j$ is stochastically monotone if $Pr\{\tilde{s}_{n+1}^j \leq y | \tilde{s}_n^j = x\}$ is non-increasing in x for every fixed y.

Theorem 15 (State-Dependent Base-Stock Policy) *Suppose the demand random variables are non-negative integers. There exists double thresholds* $\underline{u}^f$ *and* $\bar{u}^f$ *on the backup capacity investment fee,* u^f*, such that if* $u^f \leq \underline{u}^f$ *or* $u^f \geq \bar{u}^f$*, then for each*

product:

(i) A state-dependent base-stock policy is optimal.

(ii) The state-dependent base-stock level of product $j \in N$ depends on the threat level of supplier j, but not other suppliers' threat levels or other products' inventory levels.

(iii) A single-sourcing policy is optimal: for each product and in each period, the firm should procure either from the backup supplier or from the primary supplier, but not both.

(iv) If $\tilde{\mathbf{W}}^j$ is monotone, the time-stationary but state-dependent base-stock level of product j is non-decreasing in the threat level of supplier j.

Notice that the above results on the optimality of a state-dependent base-stock policy may not necessarily hold when u^f is in a middle range. This is due to the fact that the limited backup capacity needs to be *rationed* among different products, and the rationing may depend on the threat level of the suppliers.

We next provide insights into the following question.

Question 10 How much capacity in the backup system is enough?

To answer the above question, we consider the objective function of program (6.2) and establish the following lemma.

Lemma 8 *The infinite-horizon cost function (which the backup investment cost), $J_{\bar{Q}^f}(\mathbf{x}, \mathbf{s})$, is convex and non-increasing in $\bar{Q}^f$.*

The above lemma enables us to show that the optimal level of the up-front investment in the backup capacity can be easily computed using only the first order condition.

Proposition 16 *The optimal investment level, $\bar{Q}^{f*}$, is the solution to the first-order-condition: $\frac{\partial}{\partial \bar{Q}^f}\{g(u^f, \bar{Q}^f) + J_{\bar{Q}^f}(\mathbf{x}, \mathbf{s})\}|_{\bar{Q}^{f*}} = 0$*

Notice that $J_{\bar{Q}^f}(\mathbf{x}, \mathbf{s})$ cannot usually be derived analytically and needs to be numerically computed for different values of investment level, $\bar{Q}^f$, using the Bellman Eq. (6.3). However, the above result significantly reduces the computational effort required to find the optimal capacity investment level, $\bar{Q}^{f*}$, in program (6.2). After an appropriate discretization of the possible values of $\bar{Q}^f$, the search can be stopped the first time the benefit of going to the next investment level is negative, or bisection search can be used with guaranteed optimality gap.

The next result provides insights into the value of the flexible backup supplier for environments in which all products are similar with respect to the modeled parameters (i.e., a fully symmetrical case). Firms that procure relatively more products benefit more from establishing the backup supplier, but the marginal benefit diminishes as the number of products increases.

Proposition 17 (Diminishing Rate of Return) *The value of the flexible backup supplier has a diminishing rate of return (increasing concave) in n in the case of complete symmetry.*

We now provide further insights into environments that enhance the value of the flexible backup supplier.

Question 11 What characteristics of a firm make establishing a flexible backup system relatively more valuable?

A partial answer is provided in the following proposition: high procurement costs from the dedicated/primary suppliers or low procurement costs (variable or fixed/up-front investment) from the flexible supplier.

Proposition 18 (Backup Flexibility Attractiveness) *The investment in the backup capacity, $\bar{Q}^{f*}$, and the value of the backup flexible resource are increasing in vector $\mathbf{c}$, but decreasing in vector $\mathbf{c}^f$ and in u^f.*

To provide more insights into Question 11, we first need to define the following notion of reliability dominance.

Definition 5 (Reliability Dominance) *We say the t.p.m. set W (weakly) dominates the t.p.m. set W' in reliability, and denote it by $W' \leq_R W$, if $Pr\{\mathbf{s}^j(t) = 0|\mathbf{s}^j(0)\}$ for all $j \in N$ and $t \in \mathbb{Z}_+$ under W is not greater than that under W'.*

Using this definition, we can show that a firm with less reliable supply system (a) would optimally establish the backup flexible supplier at a greater (or equal) capacity, and (b) benefits more from establishing such a backup capacity:

Proposition 19 *If $W' \leq_R W$, then (i) $\bar{Q}_W^{f*} \leq \bar{Q}_{W'}^{f*}$, and (ii) $\Delta_W^f \leq \Delta_{W'}^f$.*

Furthermore, we can answer the following question:

Question 12 Can establishing a secondary flexible resource be regarded as a mechanism to compensate for the risk of disruptions in the primary suppliers?

The following result answers the above question as a qualified yes by showing that there exists a range of supply reliability improvements over which a greater or equal benefit can be achieved only by establishing a backup flexible resource (compared to only improving the reliability).

Proposition 20 *Fix all cost parameters and consider a system working under reliability governed by W' and assume there is no established flexible backup resource. There exists a capacity establishment level $\bar{Q}^f$, and a redesigned system with improved reliability W, $W' \leq_R W$, such that establishing a backup capacity of $\bar{Q}^f$ under the current system is (weakly) better than the redesigned system which has an improved reliability but no established backup capacity.*

6.4.2 No Information on Disruption Risk Levels

To model the current lack of visibility in supply chains, we now consider the case where the firm cannot observe the risk levels of its unreliable suppliers. The firm, however, may have its own belief. Here, we develop a POMDP to capture the trade-off between (a) carrying inventory of different products, (b) investing in the secondary flexible supplier, and (c) obtaining disruption information. Let $\mathbf{\Theta}(t) = (\theta^j(t), j \in N)$ be the perceived reliability *distribution* of the firm about its unreliable suppliers' disruption risk levels at period t, where $\theta^j(t) = (\theta^j_l(t), l \in S^j)$, and $\theta^j_l(t)$ is the firm's perceived probability that supplier j is in threat level l at period t satisfying $\sum_{l\in S^j} \theta^j_l(t) = 1$ ($\forall t \in \mathbb{Z}^+$, $\forall j \in N$). Then $(\mathbf{x}(t), \mathbf{\Theta}(t))$ can be considered as the state (or *information state* to be more precise) of the system at time t.

Now, let $\tilde{\mathbf{V}}^*_{(P)}(\mathbf{x}(0), \mathbf{\Theta}(0))$ represent the optimal expected total infinite-horizon discounted cost (including contracting at $t = 0$, inventory, and shortage costs) as *perceived* by the firm, if it starts with an inventory vector $\mathbf{x}(0)$ and reliability belief distribution $\mathbf{\Theta}(0)$. This value can be computed by the following program:

$$\tilde{\mathbf{V}}^*_{(P)}(\mathbf{x}(0), \mathbf{\Theta}(0)) = \min_{\bar{Q}^f \geq 0} \; g(u^f, \bar{Q}^f) + \mathbf{V}^*_{(P)\,\bar{Q}^f}(\mathbf{x}(0), \mathbf{\Theta}(0))\,, \tag{6.5}$$

where $\mathbf{V}^*_{(P)\,\bar{Q}^f}(\cdot,\cdot)$ denotes the firm's *perceived* expected infinite-horizon discounted cost given a established capacity of $\bar{Q}^f$ at time $t = 0$. This term (i.e., the second term in (6.5)) can be obtained using a POMDP model with the following optimality equation written for an arbitrary period $t \in \mathbb{Z}^+$: $\mathbf{V}^*_{(P)\,\bar{Q}^f}(\mathbf{x}(t), \mathbf{\Theta}(t)) =$

$$\begin{aligned}\min_{\mathbf{q}(t),\mathbf{q}^f(t)\geq \mathbf{0}\,:\,\sum_{k\in N} q^f_k(t) \leq \bar{Q}^f} \Big\{ & \sum_{j\in N} \big[c^f_j q^f_j + (1-\theta^j_0(t))(c^j q^j(t) + G_j(x_j(t) + q^j(t) + q^f_j(t))) \\ & + \theta^j_0(t) G_j(x_j(t) + q^f_j(t))\big] \\ & + \beta\, \mathbb{E}_{\mathbf{D}(t)} \mathbb{E}_{\mathbf{\Theta}(t+1)} \left[\mathbf{V}^*_{(P)\,\bar{Q}^f}(\mathbf{x}(t+1), \mathbf{\Theta}(t+1)) | \mathbf{x}(t), \mathbf{\Theta}(t)\right] \Big\},\end{aligned} \tag{6.6}$$

where the inventory transition rule is the same as before.

In the last term of (6.6), the belief transition rule can be any Markovian updating rule that models a firm's behavior in updating its belief. Solving (6.6) can be very difficult, because the state space (certainly the space of $\mathbf{\Theta}(t)$) is continuous. To provide a general solution method, we present the following result which significantly facilitates solving program (6.6) when the inventory state space is finite (which is not very restrictive given that numerical computation requires a finite, possibly truncated inventory state space anyway).

Proposition 21 (Smallwood and Sondik [131]) *If the inventory state space is finite with elements* $x_1, \cdots, x_k$, *define* $b_i = \mathbb{1}_{\{x=x_i\}}$, $\mathbf{b} = (b_i, i = 1, 2, \cdots, k)$ *and extended belief distribution* $\mathbf{b}' = (\mathbf{b}, \mathbf{\Theta})$. *Then the value function in (6.6) for all periods* t, *when written in terms of* $\mathbf{b}'$, *is piece-wise linear convex, and hence can be expressed as* $V^*_{(P),\bar{Q}^f,t}(\mathbf{b}') = \max_l[\sum_{i=1}^{k+\sum_{j\in N} k^j} \alpha_i^l(t) b_i']$ *for some set of vectors* $\alpha^l(t) = (\alpha_i^l(t))_{1\times(k+\sum_{j\in N} k^j)}$ *and* $l \in \mathbb{N}$, *where* b_i' *are the elements of* $\mathbf{b}'$.

The above result boils the Program (6.5)-(6.6) down to a discrete optimization setting, allows one to compute costs as perceived by the firm, and provides the firm's *perceived optimal* actions. However, to evaluate the true value of the flexible supplier and the information, we need to compute the firm's *true* cost of these actions. Perception, actions taken based on an imperfect perception, and the true realizations of Nature must be interwoven to correctly evaluate this. Let $\bar{Q}^{f*}$ and Γ^* respectively represent the firm's perceived optimal established capacity level and ordering policy derived from program (6.5)-(6.6). Then the firm's true total expected infinite-horizon discounted cost based on policy Γ^* and capacity level $\bar{Q}^{f*}$ is:

$$\tilde{\mathbf{V}}^{\Gamma^*}_{(T)\bar{Q}^{f*}}(\mathbf{x}(0)) = g(u^f, \bar{Q}^{f*}) + \mathbf{V}^{\Gamma^*}_{(T)\,\bar{Q}^{f*}}(\mathbf{x}(0)). \tag{6.7}$$

Notice that $\tilde{\mathbf{V}}^{\Gamma^*}_{(T)\,\bar{Q}^{f*}}(\cdot)$ defined above is a *policy evaluation function* based on the firm's perceived optimal contracting level ($\bar{Q}^{f*}$) and ordering policy Γ^*, which determines its *true* expected cost. To compute the second term in (6.7), let $P^j(t) =$

$(\mathcal{P}_0^j(t), \mathcal{P}_1^j(t), \cdots, \mathcal{P}_{k^j}^j(t))$, where $\mathcal{P}_l^j(t)$ denotes the true probability that the threat level of dedicated supplier j is l at period t. Because the disruption risk levels of unreliable supplier j ($j = 1, 2$) follow a Markov chain, $P^j(t)$ can be computed using its t-step transition probability matrix (t.p.m.), denoted by $(\mathbf{W}^j)^t$, given the initial risk level of this supplier. Also, let $\mathbf{q}^*(\mathbf{x}(t), \mathbf{\Theta}(t)) = \big(q^{j*}(\mathbf{x}(t), \mathbf{\Theta}(t)), j \in N\big)$ and $\mathbf{q}^{f*}(\mathbf{x}(t), \mathbf{\Theta}(t)) = \big(q_j^{f*}(\mathbf{x}(t), \mathbf{\Theta}(t)), j \in N\big)$ denote the perceived optimal ordering quantities of the firm (from the dedicated suppliers and the flexible one) under Γ^* for state $(\mathbf{x}(t), \mathbf{\Theta}(t))$ at period t. Moreover, assume $Z_j^* = x_j(t) + q^{j*}(\mathbf{x}(t), \mathbf{\Theta}(t)) + q_j^{f*}(\mathbf{x}(t), \mathbf{\Theta}(t))$ and $Y_j^* = x_j(t) + q_j^{f*}(\mathbf{x}(t), \mathbf{\Theta}(t))$ show the firm's possible inventory of product j right after ordering at period t (where for simplicity we have removed indexes showing the dependency of Z_j^* and Y_j^* on the state of the system). Then for an arbitrary period $t \in \mathbb{Z}^+$:$\mathbf{V}_{(T)\,\bar{Q}^{f*}}^{\Gamma^*}(\mathbf{x}(t)) =$

$$\sum_{j\in N} \left[c_j^f\, q_j^{f*} + (1 - \mathcal{P}_0^j(t))\,(c^j q^{j*} + G_j(Z_j^*)\,) + \mathcal{P}_0^j(t)\, G_j(Y_j^*)\,\right] \tag{6.8}$$

$$+ \beta \sum_{m_j\in\{0,1\}, j\in N} \Big[\prod_{j\in N} \left[(\mathcal{P}_0^j(t))^{m_j}(1 - \mathcal{P}_0^j(t))^{1-m_j}\right] \times \oint_0^\infty V_{(T)\,\bar{Q}^{f*}}^{\Gamma^*}(m_j Y_j^* + (1 - m_j)Z_j^* - \xi_j, j \in N) \prod_{j\in N} dF_j(\xi_j)\Big],$$

where $\oint_0^\infty$ represents the n-fold Riemann-Stieltjes integral: $\oint_0^\infty = \underbrace{\int_0^\infty \cdots \int_0^\infty}_{n}$.

Now, we are able to compute the Improvement Percentage (IP) in the firm's true expected total discounted cost due to establishing the flexible backup resource ($IP_{(F)}\%$) in a no-information infinite-horizon setting. For simplicity, we assume that $\mathbf{x}(0) = \mathbf{0}$, i.e., there is no initial on-hand inventory. Without information, $IP_{(F)}\%$, a measure for the value of the flexible supplier, can be computed by:

$$IP_{(F)}\% = \frac{\mathbf{V}_{(T)\,\bar{Q}^f=0}^{\Gamma^{'*}}(\mathbf{0}) - \tilde{\mathbf{V}}_{(T)\,\bar{Q}^{f*}}^{\Gamma^*}(\mathbf{0})}{|\mathbf{V}_{(T)\,\bar{Q}^f=0}^{\Gamma^{'*}}(\mathbf{0})|} \times 100\,, \tag{6.9}$$

where $\acute{\Gamma}^*$ denotes the firm's optimal perceived ordering policy in the absence of the flexible supplier, i.e., when $\bar{Q}^f = 0$.

We illustrate these concepts and generate some managerial insights through a numerical study. As mentioned in the model formulation, it is more numerically tractable to use the lost sales variant of the model, eliminating states with a backlog. Moreover, if we include a revenue of r_j per unit of type j sold, then the behavior of the lost sales model becomes similar to that of the original model. Hence, we use the lost sales variant of the model for the remainder of the chapter. It should be noted that the computational methodology developed above carries over to this variant.

Study 1 (No-Information) Consider the parameter settings presented in Table 6.4 of Appendix B for different firms in a market with two underlying products where demand for both products (after possible discretization and scaling) follows a discrete uniform distribution on [1; 5]. Settings 1-4 in Table 6.4 differ only in the reliability beliefs. Settings 1-4 progressively increase from low reliability to high. Settings 5-8 include variations in the other parameters as well (see Appendix B for more details). Here, for clarity of insights, we assume (1) the capacity investment is linear in the capacity, (2) c_j^f $(j = 1, 2)$ is negligible compared to u^f (we will drop this assumption later in study 7), and (3) $u^f = \tilde{u}^f/(1-\beta)$ where $\tilde{u}^f$ (presented in Table 6.4) denotes the per period average cost of contracting with the flexible supplier so that the value $\tilde{u}^f$ in a single-period setting is equivalent to u^f in this study. We also assume firms do not change their beliefs over time; otherwise, the value of the flexible backup capacity and the value of information will depend on the updating behavior implemented. Moreover, we consider four different disruption risk transition probability matrices (t.p.m.'s) as presented in Appendix B part (ii) for each parameter setting of Table 6.4. Table 6.1 illustrates the computational results for different parameter and t.p.m. settings (32 different settings in all) with a discount factor of $\beta = 0.9$ as well as the related errors in the firms' reliability beliefs. For instance, a firm with parameter setting 6 and unreliable suppliers with risk level dynamics based on t.p.m. setting

3 has a reliability belief error of $\Upsilon = (-0.070, 0.085)$ and will choose to contract with (or establish) the flexible supplier for $\bar{Q}^{f*} = 4$ units (out of a scaled maximum demand of 5). Consequently, it can decrease its true infinite-horizon discounted costs by $IP_{(F)}\% = 16.34\%$ through contracting with a flexible backup supplier. In fact, this firm observes contracting to be valuable and is lucky that it is also truly valuable. However, one may note that implementing a backup flexible supplier is not always valuable. As an example, a firm with parameter setting 1 and with t.p.m. setting 3 perceives establishing (or contracting with) the flexible supplier to be valuable and hence forms a contract for 3 (scaled) units. However, as Table 6.1 shows, this decision not only does not reduce its cost, but will even increase it by 42.12% (i.e., $IP_{(F)}\% = -42.12\%$). On the other hand, for a firm with parameter setting 8 and under t.p.m. setting 4, a capacity reservation contract with a flexible supplier is a strong mitigation technique that can result in a 152.50% total cost reduction (which means that the cost of the current system is replaced with a profit roughly half as large). The firms with parameter settings 2-4 (because of erroneous reliability belief) do not perceive establishing the backup capacity to be valuable and hence cannot benefit from a flexible supplier, even if they have the option to do so. Considering that currently most firms do not monitor their suppliers carefully, and hence, there is a lack of disruption risk information in most supply chains, we gain two important observations from the this study:

Observation 19 Firms that do not monitor their suppliers on a regular basis should be careful to analyze the opportunity to contract with (or invest in) a backup flexible supplier, and this can be done using sensitivity analysis.

Observation 20 Implementing flexible backup supply capacity can be a potent way to mitigate a lack of disruption information in supply chains.

Table 6.1: Numerical Study 1 (No-Information.) [ϵ^j: firm's reliability perception error of unreliable supplier j].

Setting No.	t.p.m. Setting	ϵ^1	ϵ^2	$\bar{Q}^{f*}$	Perceived Cost ($\bar{Q}^{f*}$)	True Cost ($\bar{Q}^{f*}$)	Perceived Cost ($\bar{Q}^{f}=0$)	True Cost ($\bar{Q}^{f}=0$)	$IP_{(F)}\%$
1	1	-0.015	-0.015	3	-61.18	-62.58	-54.06	-57.23	9.34
	2	-0.015	-0.120			-64.60		284.91	122.69
	3	-0.120	-0.015			404.91		284.91	-42.12
	4	-0.120	-0.120			404.91		284.91	-42.12
2	1	0.035	0.085	0	-71.23	-57.52	-71.23	-57.52	0.00
	2	0.035	-0.020			-68.39		-68.39	0.00
	3	-0.070	0.085			-70.80		-70.80	0.00
	4	-0.070	-0.020			284.91		284.91	0.00
3	1	0.085	0.035	0	-72.38	-57.52	-72.38	-57.52	0.00
	2	0.085	-0.070			-68.39		-68.39	0.00
	3	-0.020	0.035			-70.80		-70.80	0.00
	4	-0.020	-0.070			284.91		284.91	0.00
4	1	0.135	0.135	0	-88.12	-57.52	-88.12	-57.52	0.00
	2	0.135	0.030			-68.39		-68.39	0.00
	3	0.030	0.135			-70.80		-70.80	0.00
	4	0.030	0.030			-81.65		-81.65	0.00
5	1	0.035	0.085	4	-769.32	-760.15	-750.00	-720.16	5.55
	2	0.035	-0.020			-767.60		-748.00	2.62
	3	-0.070	0.085			-769.46		-737.42	4.34
	4	-0.070	-0.020			-776.24		284.91	372.45
6	1	0.035	0.085	4	-241.13	-235.90	-217.08	-190.29	23.97
	2	0.035	-0.020			-240.64		-215.77	11.52
	3	-0.070	0.085			-240.04		-206.32	16.34
	4	-0.070	-0.020			-244.39		389.91	162.68
7	1	0.135	0.085	1	-244.16	-208.37	-243.74	-196.82	5.87
	2	0.135	-0.020			-225.31		-223.36	0.87
	3	0.030	0.085			-225.77		-216.48	4.29
	4	0.030	-0.020			-242.41		-243.03	-0.26
8	1	0.035	0.035	1	-216.03	-204.87	-214.51	-200.58	2.14
	2	0.035	-0.070			-221.26		-219.49	0.81
	3	-0.070	0.035			-220.17		-220.24	-0.03
	4	-0.070	-0.070			-236.19		449.91	152.50

6.4.3 The Value of Flexibility: A Single Flexible Backup Supplier or Dedicated Ones?

In the previous sections, we examined the value of a reliable flexible secondary supplier. Here, we investigate the value of *supply flexibility* as a mechanism to mitigate supply risk, and we do so in environments with full disruption state information as well as without it. To this end, we compare the performance of two systems, (A) and (B). In addition to the existing unreliable dedicated suppliers, System (A) reserves a dedicated reliable secondary supply capacity for each product, and System (B) invests in a single reliable flexible backup supplier. The comparison of (B) relative to (A) exposes the difference that flexibility makes. Moreover, to investigate the interplay between the value of flexibility and information, we compare these two systems under both full and no-information scenarios. The analytical framework to compute the infinite-horizon perceived cost and the true cost is the same as described in Sections (6.4.2) and (6.4.1). The only difference is that for System (A) one can compute the

Table 6.2: The Value of Supply Flexibility under No-Information.

Setting No.	t.p.m. Setting	System (A): Two Inflex. Secondary $\bar{Q}_1^{f*} + \bar{Q}_2^{f*}$	Prcvd. Cost	True Cost	System (B): One Flex. Secondary $\bar{Q}^{f*}$	Prcvd. Cost	True Cost	Value of Flexibility (%)
1	1	2	-58.15	-42.78	3	-61.18	-62.5	46.28
	2			138.39			-64.60	146.68
	3			183.46			404.91	-120.71
	4			364.63			404.91	-11.05
2	5	0	-81.85	-21.71	0	-81.85	-57.59	165.27
	6			-68.4			-68.51	0.16
	7			-35.05			-70.93	102.37
	8			-81.74			-81.85	0.13
3	1	4	-762.87	-673.96	4	-769.32	-760.15	12.79
	2			-674.31			-767.60	13.83
	3			-682.32			-796.46	16.73
	4			-682.67			-776.24	13.71
4	1	4	-235.56	-206.36	4	-241.13	-235.90	14.31
	2			-171.86			-240.64	40.02
	3			-210.08			-240.04	14.26
	4			-175.58			-244.39	39.19
5	1	0	-243.74	-196.72	1	-244.16	-208.37	5.92
	2			172.54			-225.31	230.58
	3			-216.38			-225.70	4.31
	4			152.88			-242.41	258.56
6	1	0	-214.51	-200.49	1	-216.03	-204.87	2.18
	2			209.91			-221.26	205.41
	3			76.79			-220.17	386.72
	4			57.9			-236.19	507.93
Average								**87.316**

costs separately for each product since there is no connection between them. The following study investigates the value of information and flexibility, and generates insights into their interactions.

Study 2 (Flexibility and the Effect of Information) Consider a firm that procures two products from existing unreliable primary suppliers, under the parameter settings presented in Table 6.5 in Appendix B. The demand distributions and other assumptions are the same as those in Study 1. Tables 6.2 and 6.3 present the optimal cost of the firm without and with information, respectively. In these tables the value of supply flexibility is reported as the percentage improvement in the cost of the firm if it implements System (B) instead of (A). The last column of Table 6.3 also reports the pure percentage-wise value of information (i.e., information in the absence of flexibility) by comparing the true cost of System (A) with and without information. From these tables we gain several interesting strategic insights. First, a flexible backup supplier is always more valuable than two dedicated backup ones for a firm that can perfectly assess its suppliers' risk levels. In other words, implementing flexibility in a supply system cannot be harmful under perfect information, as one would probably expect. However, it may surprise many that this is not the case for

Table 6.3: The Value of Supply Flexibility with Full-Information and the Pure Value of Information (without Flexibility).

Setting No.	t.p.m. Setting	**Two Inflex. Secondary**		**One Flex. Secondary**		Value of Flexibility (%)	Pure Value of Information (%)
		$\bar{Q}_1^{f*} + \bar{Q}_2^{f*}$	Cost	$\bar{Q}^{f*}$	Cost		
1	1	1	-70.95	2	-77.77	9.61	65.85
	2	0	-77.51	1	-80.29	3.59	156.01
	3	1	-79.43	1	-81.81	3.00	143.30
	4	0	-85.99	1	-86.02	0.03	123.58
2	5	1	-67.13	2	-76.69	14.24	209.21
	6	0	-74.11	2	-79.39	7.12	8.35
	7	1	-77.98	2	-80.87	3.71	122.48
	8	0	-84.96	1	-85.48	0.61	3.94
3	1	2	-767.93	3	-786.86	2.47	13.94
	2	1	-775.8	3	-791.23	1.99	15.05
	3	2	-779.2	3	-791.80	1.62	14.20
	4	1	-787.07	2	-796.17	1.16	15.29
4	1	4	-235.34	4	-248.00	5.38	14.04
	2	3	-240.12	3	-249.91	4.08	39.72
	3	3	-239.6	3	-249.22	4.02	14.05
	4	2	-244.38	3	-251.29	2.83	39.18
5	1	1	-227.77	3	-243.53	6.92	15.78
	2	1	-241.44	2	-249.19	3.21	239.93
	3	0	-237.89	2	-247.26	3.94	9.94
	4	0	-251.56	1	-254.16	1.03	264.55
6	1	0	-225.05	2	-232.64	3.37	12.25
	2	0	-238.73	1	-241.26	1.06	213.73
	3	0	-237.88	1	-239.32	0.61	409.78
	4	0	-251.56	0	-251.56	0.00	534.47
Average						**3.566**	**112.443**

the firms with no-information. In fact, we can observe the following:

Observation 21 With the lack of disruption information, supply flexibility can be harmful (up to 120.71% in our test suite). However, on average implementing supply flexibility for firms with no-information can be regarded as a strong mitigation technique that resulted in a cost improvement of 87%, compared to the case of using dedicated backup suppliers (see Table 6.2).

Another important observation from comparing the value of flexibility columns in Tables 6.2 and 6.3 illuminates the effect of information on the benefit of flexibility.

Observation 22 Obtaining information may or may not augment the benefit of supply flexibility. On average, however, supply flexibility is much more valuable for firms that cannot monitor the risk levels of their suppliers than those with full information.

Next, it is worth considering the pure value of information and compare it with the value of flexibility:

Observation 23 In contrast with implementing flexibility, obtaining information is always valuable. The pure value of information (information without flexibility) is on average higher than the pure value of flexibility (flexibility without information), 112% vs. 87% in our test suite.

The above observation suggests that obtaining information is a better insurance against disruptions than flexibility. (This value of information without flexibility derives from the ability of the system to dynamically adjust inventory safeguards as the threat levels change.) However, considering the expensive costs (and technical difficulty) of thoroughly and dynamically monitoring suppliers (e.g. cost of establishing a threat level advisory system), one may observe that:

Observation 24 Implementing flexibility in the backup system can significantly, though not completely, compensate for the lack of disruption information in supply chains.

Finally, comparing the investment in the backup capacity with and without flexibility, we observe the following:

Observation 25 A firm (with or without information) will reserve at least as much capacity from a flexible secondary supplier than the amount reserved in total from dedicated secondary ones.

In fact, the flexibility of a supplier provides the buyer with greater benefit, justifying reserving more backup capacity because of the economic advantage of shifting the orders whenever necessary (capacity pooling).

In the next study, we further investigate the capacity pooling advantage of the flexible backup supplier in a full-information scenario and we analyze its sensitivity with respect to the linear purchasing cost term of the capacity reservation contract.

Study 3 (Capacity Pooling Advantage) Consider a firm procuring two products from unreliable suppliers under a full-information scenario where c_1^f and c_2^f are, unlike the previous studies, not negligible. Specifically, let $\tilde{u}^f = 1.0$, $c_1^f = 0.3$, $c_2^f = \Delta\, c^2$ (where Δ scales the cost of dedicated supplier 2). Assume the dynamics of disruption risks are defined by t.p.m. setting 1 (see Table 6.6 of Appendix B) and the demand distributions and other assumptions are the same as in the previous study. Fig. 2 (left) reveals the insight that the value of the flexible secondary supplier is more than the summation of benefits that can be obtained separately for each of the products through dedicated secondary suppliers. This is mainly due to the *capacity pooling advantage* of the flexible secondary supplier; when one of the primary suppliers is in a high risk threat level and the other is in a low threat level, the reserved pooled capacity can be used as needed. However, using the difference between the two curves depicted in Fig. 2 (left), we have the following observation:

Observation 26 The pooling advantage is not monotone in Δ and has its maximum effect at $\Delta = 0.5$. However, as Δ increases, the pooling advantage vanishes: the flexible supplier can only be used for product 1, performing as a dedicated supplier.

Fig.6.4.3 (right) depicts the corresponding optimal investment levels in the secondary suppliers. As this figure shows, the sum of optimal capacities required for product 1 and 2 in the case of two dedicated backup suppliers is never larger than the optimal capacity reserved with the flexible secondary supplier. This latter observation coincides with Observation 25 described in Study 2 (where the linear purchasing costs of the capacity reservation contract were assumed to be negligible). These results highlight the effect of capacity pooling, which can be a significant benefit of flexibility.

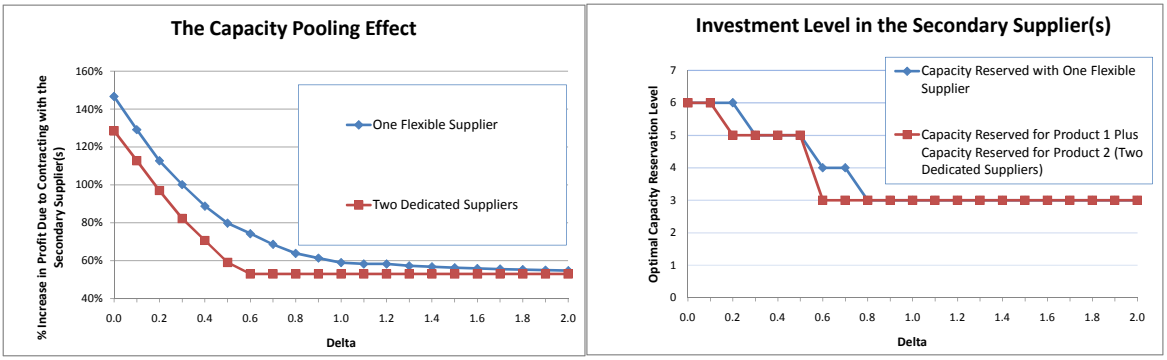


Figure 6.2: The Pooling Effect under Full Information ($\beta = 0.9, c^1 = c^2 = 2, r_1 = r_2 = 3.5, p_1 = p_2 = 2.5, h_1 = 1, h_2 = 1.2$).

6.5 Summary and Conclusion

We addressed the strategic value of two powerful supply risk mitigation mechanisms: (1) contracting with (or establishing) a secondary flexible supplier, and (2) monitoring the dynamic risk levels of primary suppliers. We modeled the dynamics of disruptions as discrete time Markov chains and considered different cases of information availability with respect to the disruption risk levels, including scenarios in which such information may or may not be possible to obtain..

We first studied how firms should update their inventory safeguards to effectively mitigate against dynamic disruption risks. Under full disruption risk information, we showed that when the investment cost in the backup supplier is not in a middle range, a state-dependent base-stock policy is optimal. Furthermore, a single sourcing policy becomes optimal for each of the products; the firm should either procure from the flexible backup supplier or from the dedicated supplier, but not both. Moreover, if the (transformed) Markov chain governing the disruption risk of a dedicated supplier is stochastically monotone, then base-stock levels are non-decreasing in the threat levels. We also showed how firms can effectively quantify the decision of the investment in the backup capacity.

In cases of symmetry across the products, we showed that the flexibility in the

backup system has a diminishing rate of return in the underlying number of products; the flexibility in the backup system is more valuable for firms that procure several products, but the marginal benefit diminishes as the number of underlying products increases. We also formally proved that investing in the backup resource is more valuable for firms that procure from relatively more expensive unreliable suppliers.

We next introduced a notion of reliability dominance and showed that a less reliable supply system will require a greater backup capacity investment, and also will result in greater benefits from such an investment. More importantly, we showed that establishing a flexible secondary capacity can effectively compensate for the the risk of disruptions in the primary suppliers.

We then studied the case where the firm cannot monitor its unreliable suppliers: a no disruption information case. We model this case as a POMDP, differentiated between perception and reality, and developed a methodology for computing the firm's true cost of its procurement strategy and imperfect perception of system state. Using this framework, we developed measures to quantify (a) the value of investing in a dedicated (inflexible) backup supplier for each product, (b) the value of a single flexible backup supplier, and (c) the value of disruption information with or without a flexible backup supplier.

We numerically used these measures and gained five main managerial insights: (1) Benefiting from a flexible backup supplier can be a potent mitigation mechanism against the current lack of disruption risk information in supply chains. (2) Although implementing flexibility in supply system is always valuable under perfect information, firms should be aware that supply flexibility can be even harmful when disruption risk information is incomplete. (3) Obtaining information may or may not further augment the benefit of supply flexibility. However, on average, implementing supply flexibility is more valuable for the firms that cannot thoroughly monitor the risk levels of their suppliers than those that can. (4) In contrast with implementing flexibility, obtaining information is always purely valuable. Moreover, the pure value

of information (information without flexibility) is on average higher than the pure value of flexibility (flexibility without information). This suggests that information is on average a better "insurance" against potential disruptions compared to supply flexibility. However, considering the expensive costs (and intrinsic technical difficulties) of thoroughly and dynamically monitoring suppliers, our observations strongly suggest that supply flexibility is a potent substitute for the need of obtaining information on dynamic risk levels of suppliers in supply chains. (5) A firm (with or without info) will reserve at least as much capacity from a secondary flexible supplier as the amount reserved in total from two dedicated secondary ones. Indeed, the flexibility of a supplier provides the buyer with greater benefit, justifying reserving more backup capacity because of the economic advantage of shifting the orders whenever necessary (capacity pooling). This latter observation also justifies charging a higher cost by flexible suppliers for reserving their flexible capacity.

The analyses, framework, and insights presented in this chapter can guide new practices to effectively increase the resilience of supply chains, especially in an era when supply chains face grave disruption risks. Increasing resilience in supply chains can in turn enable firms to deliver products with better availability and better prices to end customers, yielding social benefits. While in this study we focused on the cost of a firm, future research may examine the possibility of creating such broader social advantages.

6.6 Appendix A: Proofs

Proof of Theorem 15. (i) We first consider the case with $u^f = 0$ (i.e., when the backup capacity is unlimited) and show that a *state-dependent* base-stock policy is optimal. For the ease of notation let $J_{\bar{Q}^f} = J_{\bar{Q}^f}(\mathbf{x}(0), \mathbf{s}(0))$, for any reserved capacity $\bar{Q}^f$. Notice that by assumption $c_j^f + u^f > 0$ for all $j \in N$, and hence, $c_j^f > 0$ (for all $j \in N$) as $u^f = 0$. Thus, there exists a finite capacity investment level $M < \infty$ for some

large enough integer M, such that the total optimal cost with $Q^f = M$ and $Q^f = \infty$ are equal. Since M is not necessarily unique, choose the smallest such M and denote it by M^*. M^* is then such that $J_{M^*-1} > J_{M^*}$, and hence, there exists $\epsilon > 0$ such that $J_{M^*-1} - J_{M*} = \epsilon$. To show that a state-dependent base-stock policy is optimal for each product $j \in N$ when $u^f = 0$, consider the following analog inventory control problem with a state-dependent ordering cost (and stationary demand). Suppose the firm procures each product j from a *single* supplier where the ordering cost is state-dependent and given by $c'_j(s^j) = \min\{c_j^f, c^j\}\mathbb{1}_{\{s^j>0\}} + c_j^f\mathbb{1}_{\{s^j=0\}}$. Next notice that this inventory problem is a special case of the model considered in Section 9 of Song and Zipkin [133]. Following the result provided in Song and Zipkin [133] for a single product setting, a state-dependent base-stock policy is optimal for each product $j \in N$ when $u^f = 0$. Next, define $\Delta(u^f) = g(u^f, M^*) - g(u^f, M^* - 1)$. Notice that since with $u^f = 0$, M^* is an optimal capacity investment level but $M^* - 1$ is not, we have $\Delta(0) < \epsilon = J_{M^*-1} - J_{M^*}$. Thus, since $\Delta(u^f)$ is increasing in u^f (as $g(u^f, \bar{Q}^f)$ is supermodular by assumption), define the threshold $\bar{u}^f = \sup u^f \geq 0 : \Delta(u^f) < \epsilon$ (with the convention that $\sup \emptyset = 0$). Next observe that for all $u^f \leq \bar{u}^f$, the optimal capacity level is still M^*. That is for all $u^f \leq \bar{u}^f$ the optimal capacity investment level is the same as that with $u^f = 0$. Thus, the optimization problem (6.3) with $u^f = 0$ and any $u^f < \bar{u}^f$ is the same, and hence, the optimal control policy for any $u^f < \bar{u}^f$ is the same state-dependent base stock policy that is optimal with $u^f = 0$. Next, define $\Delta'(u^f) = g(u^f, 1) - g(u^f, 0)$, since $\Delta'(u^f)$ is increasing, let $\overline{u}^f = \inf u^f : \Delta'(u^f) > J_0 - J_1$ (with the convention that $\inf \emptyset = \infty$), and notice that for all $u^f \geq \overline{u}^f$ it is optimal to set the optimal investment level to $\bar{Q}^{f*} = 0$ (since the cost of investing in $\bar{Q}^f = 1$ compered to $\bar{Q}^f = 0$, Δ' is greater than its benefit, $J_0 - J_1$). This completes the proof of part (i). The proof of part (ii) directly follows from the proof of part (i), since it shows that when $u^f \leq \underline{u}^f$ or $u^f \geq \overline{u}^f$, the system decomposes to $|N|$ single-product settings. The proof of part (iii) is trivial for the case with $u^f \geq \overline{u}^f$, since it is optimal to set $\bar{Q}^f = 0$ from the proof of part (i). For

the case where $u^f \leq \underline{u}^f$ notice that as discussed in the proof of part (i), for each product j, the system is equivalent to a system with a single procurement source but with state dependent cost $c'_j(s^j) = \min\{c_j^f, c^j\}\mathbb{1}_{\{s^j>0\}} + c_j^f \mathbb{1}_{\{s^j=0\}}$. This shows that when the unreliable supplier is up, the firm only orders from the source with the minimum purchasing cost, and orders from the backup capacity, otherwise. To prove part (iv), first consider the case where $u^f \leq \underline{u}^f$. From above, it follows that the system transforms to a collection of single-product systems, where the purchasing cost of product j is state-dependent and given by $c'_j(s^j) = \min\{c_j^f, c^j\}\mathbb{1}_{\{s^j>0\}} + c_j^f \mathbb{1}_{\{s^j=0\}}$. The result then follow from Theorem 4 of Song and Zipkin [133]. Similarly, with $u^f \geq \overline{u}^f$, the result for each product follows from Theorem 4 of Song and Zipkin [133]. □

Proof of Lemma 8. To show convexity of function $J_{\bar{Q}^f}(\cdot,\cdot)$ in $\bar{Q}^f$, first consider a finite horizon version of (6.3), and let $J_{\bar{Q}^f,n}(\cdot,\cdot)$ denote the cost when there are n periods to go. Since $J_{\bar{Q}^f,0}(\cdot,\cdot) = 0$, using induction, suppose $J_{\bar{Q}^f,n}(\cdot,\cdot)$ is convex in $\bar{Q}^f$. To observe $J_{\bar{Q}^f,n+1}(\cdot,\cdot)$ is then also convex in $\bar{Q}^f$, notice that since the set $\{\mathbf{q}(t), \mathbf{q}^f(t) \geq 0 : \sum_{j\in N} q_j^f(t) \leq \bar{Q}^f\}$ for any $\bar{Q}^f$, and functions $G_j(\cdot)$ are convex, from the finite horizon version of (6.3), $J_{\bar{Q}^f,n+1}(\cdot,\cdot)$ is a minimization of a convex function over a convex set. Hence, $J_{\bar{Q}^f,n+1}(\cdot,\cdot)$ is convex in $\bar{Q}^f$ by the preservation of convexity under minimization (see , e.g., Lemma A.4. page 227 of Porteus 2002). Therefore, $J_{\bar{Q}^f,n}(\cdot,\cdot)$ is convex in $\bar{Q}^f$ for any n. It follows that the infinite-horizon cost, $J_{\bar{Q}^f}(\cdot,\cdot)$, is also convex in $\bar{Q}^f$. Moreover, $J_{\bar{Q}^f}(\cdot,\cdot)$ is also non-increasing in $\bar{Q}^f$, since increasing $\bar{Q}^f$ will only enlarge the feasible set of the minimum operator (6.3). □

Proof of Proposition 16. Notice that the objective function of program (6.2) is convex in $\bar{Q}^f$, since (1) $g(u^f, \bar{Q}^f)$ is convex in $\bar{Q}^f$ by assumption, and (2) $J_{\bar{Q}^f}(\cdot,\cdot)$ is convex in $\bar{Q}^f$ by Lemma 8. Hence, the fist order condition characterizes the optimal investment level, $\bar{Q}^{f*}$. □

Proof of Proposition 17. Fix the initial state, suppose $\bar{Q}^{f*} = \psi(n)$ for some function ψ, and let $\tilde{J}^n(\bar{Q}^f) = g(u^f, \bar{Q}^f) + J^n(\bar{Q}^f)$ denote the total optimal cost when

the backup capacity is $\bar{Q}^f$ (with its minimizer $\bar{Q}^{f*}$) and there are n (symmetric) products. We first show the concavity property of the value of the flexible backup resource. Since the value of the backup resource is $\Delta^{f,n} = \tilde{J}^n(0) - \tilde{J}^n(\psi(n))$, and $\tilde{J}^n(0)$ is trivially linear in n (due to decomposition of the system into n equivalent systems when $\bar{Q}^f = 0$), to show that $\Delta^{f,n}$ is concave in n, it is sufficient to show that $\tilde{J}^n(\psi(n)) = g(u^f, \psi(n)) + J^n(\psi(n))$ is convex in n. To this end, it is convenient to take n as a continuous variable (extending the definition in the case of complete symmetry), and show that $\partial J^n(\psi(n))/\partial n \geq 0$ (notice that for a function f, if $f(x)$ is convex for $x \in \mathbb{R}_+$, then $f(n)$ is convex in $n \in \mathbb{N}$). Since $g(u^f, \bar{Q}^f)$ is convex increasing in $\bar{Q}^f$ by assumption, and $\psi(n)$ is convex increasing in n (due to the capacity pooling effect), using the chain rule it can be seen that $\partial g(u^f, \psi(n))/\partial n \geq 0$. Thus, it is sufficient to show that $J^n(\psi(n))$ is convex in n, where $J^n(\psi(n))$ is defined by (6.3). Convexity of $J^n(\psi(n))$ in n follows the preservation of convexity under minimization (see , e.g., Lemma A.4. page 227 of Porteus 2002). To see this, consider the finite horizon version of (6.3), and similar to the proof of Proposition 8 use induction on the number of periods to go. Trivially, when there is $k = 0$ periods to go $J_k^n(\psi(n)) = 0$, and hence, $J_k^n(\psi(n))$ in convex in n. Next suppose $J_k^n(\psi(n))$ is convex in n for some number of periods to go, $k \in \mathbb{N}$. To show that $J_{k+1}^n(\psi(n))$ is also convex in n, observe that $J_{k+1}^n(\psi(n))$ can be written as a $\min_{\mathbf{y} \in \mathbf{Y}(n)} f(n, \mathbf{y})$ for some convex function f, vector $\mathbf{y}$, and convex set $y \in \mathbf{Y}(n)$ (i.e., minimum of a convex function over a convex set). Hence, $J_{k+1}^n(\psi(n))$, is convex in n, and therefore, $J_k^n(\psi(n))$ is convex in n for any number of periods to go, k. It then follows that the infinite-horizon cost function, $J^n(\psi(n))$ is convex. To show that $\Delta^{f,n} = \tilde{J}^n(0) - \tilde{J}^n(\psi(n))$ is increasing in n, notice that $\tilde{J}^n(0)$ is linear in n but $\tilde{J}^n(\psi(n))$ is less than linear in n due to the well-known capacity pooling effect. Hence, it can be shown that $\Delta^{f,n}$ is increasing in n which completes the proof. □

Proof of Proposition 18. Fix the state and consider From Bellman Eq. (6.3). Notice that increasing c_j will decrease the optimal q_j^*. Next let $w_j = x_j + q_j^* + q_j^{f*}$ be

the optimizer of $G_j(x_j + q_j + q_j^f)$. With a a lower q_j^*, w_j is achieved with a higher q_j^{f*}. This will in turn require a (weakly) greater capacity investment level, $\bar{Q}^{f*}$. Thus, $\bar{Q}^{f*}$ is increasing in $\mathbf{c}$. This also results in a (weakly) greater benefit due to establishing the backup capacity (Δ^f), i.e., comparing the cost with $\bar{Q}^{f*}$ and that with $\bar{Q}^f = 0$. Using a similar argument, notice that increasing c^f will decrease $\bar{Q}^{f*}$ and Δ^f. To see the effect of u^f, consider program (6.2). Since by assumption $g(u^f, \bar{Q}^f)$ is supermodular, increasing u^f will decrease $\bar{Q}^{f*}$. It then follows that (Δ^f) is also decreasing in u^f. $\square$

Proof of Proposition 19. From Bellman Eq. (6.3) and using the envelope theorem, it can be seen that, at any capacity $\bar{Q}^f$, $\partial J^W/\partial \bar{Q}^f \leq \partial J^{W'}/\partial \bar{Q}^f$. That is the cost reduction due to adding backup capacity (excluding the investment cost) is higher under the less reliable system. Also, from Proposition 16, the optimal capacity investment levels $\bar{Q}_W^f$ and $\bar{Q}_{W'}^f$ are the solutions to $\partial J^W/\partial \bar{Q}^f = k$ and $\partial J^{W'}/\partial \bar{Q}^f = k$ (where $k = -\partial g(u^f, \bar{Q}^f)/\partial \bar{Q}^f$), respectively. Thus, the proof of part (i) is complete, since $\partial J^W/\partial \bar{Q}^f$ and $\partial J^{W'}/\partial \bar{Q}^f$ are both non-decreasing functions of $\bar{Q}^f$ (since J is convex in $\bar{Q}^f$ by Lemma 8). To prove part (ii), fixing the initial state, let $J^W(\bar{Q}^f)$ and $J^{W'}(\bar{Q}^f)$ denote the operational cost (i.e, not including the capacity investment cost) when a backup capacity of $\bar{Q}^f$ is established under radiabilities governed by W and W', respectively. Also, let $\bar{Q}_W^f$ and $\bar{Q}_{W'}^f$ be the optimizers of Program (6.2) under W and W', respectively. Since $\bar{Q}_{W'}^f$ is the optimizer of Program (6.2) under W', we have :

$$J^{W'}(\bar{Q}_{W'}^f) + g(u^f, \bar{Q}_{W'}^f) \leq J^{W'}(\bar{Q}_W^f) + g(u^f, \bar{Q}_W^f). \tag{6.10}$$

Furthermore, since the difference in the operational costs is maximum when there is no backup capacity (as backup capacity alleviates the impact of unreliability), comparing the zero backup capacity with that of establishing a backup capacity of $\bar{Q}_W^f$ for both systems, we have:

$$J^W(0) - J^{W'}(0) \leq J^W(\bar{Q}_W^f) - J^{W'}(\bar{Q}_W^f). \tag{6.11}$$

Adding inequalities (6.10) and (6.11) results in $\Delta_W^f \leq \Delta_{W'}^f$, which completes the

proof. □

Proof of Proposition 20. From Bellman Eq. (6.3), it is easy to show that for any fixed state and capacity investment $\bar{Q}^f$, we have $J_{\bar{Q}^f} W(\cdot,\cdot) \leq J^{W'}_{\bar{Q}^f}(\cdot,\cdot)$. That is, fixing the state and capacity level, the cost under the reliability governed by W is not greater than that under the reliability governed by W'. Then, if $\bar{Q}^{f*}_{W'}$ denotes the optimal backup capacity level under W', choose $W(W' \leq_R W)$ such that $J_0 W(\mathbf{x}(0),\mathbf{s}(0)) \in [g(u^f,\bar{Q}^{f*}_{W'}) + J^{W'}_{\bar{Q}^{f*}}(\mathbf{x}(0),\mathbf{s}(0)),\ g(u^f,0) + J_0^{W'}(\mathbf{x}(0),\mathbf{s}(0))]$. Since $g(u^f,\bar{Q}^f)+J^{W'}_{\bar{Q}^f}(\mathbf{x}(0),\mathbf{s}(0))$ is continuous in $\bar{Q}^f$, there exist a $\bar{Q}^{f1}$ such that $J_0^W(\mathbf{x}(0),\mathbf{s}(0)) = g(u^f,\bar{Q}^{f1}) + J^{W'}_{\bar{Q}^{f1}}(\mathbf{x}(0),\mathbf{s}(0))$. Next, since $g(u^f,\bar{Q}^f) + J^{W'}_{\bar{Q}^f}(\mathbf{x}(0),\mathbf{s}(0))$ is (convex by Lemma 1 and) decreasing in $\bar{Q}^f$ in interval $[\bar{Q}^{f1},\bar{Q}^{f*}_{W'}]$, choosing any $\bar{Q}^f \in [\bar{Q}^{f1},\bar{Q}^{f*}_{W'}]$ together with W' satisfies the result of the proposition. □

Proof of Proposition 21. The proof directly follows that of Smallwood and Sondik (1973) with a small change of notation. □

6.7 Appendix B: Parameter Settings

(i) Suite of Parameter Settings Considered in Numerical Studies

The parameter settings considered are as follow. The first four settings are identical except for the beliefs. The other settings include variations on other parameters as well. It is noteworthy that in all tables of this appendix, $\tilde{u}^f$ represents the *average cost per period* per unit of capacity reservation cost with the flexible supplier.

The parameter settings considered for Study 2 are presented in Table 6.5, which builds upon Table 6.4. The first two settings (similar to settings 1-4 of Table 6.4) represent same firms but with different beliefs. It is noteworthy that, for instance, Setting 2 together with t.p.m. setting 8 of Table 6.6 represents a special case where the firm knows the steady state distribution of reliability of its supplier. However, in real-world firms (in no-information environments) do not know the steady state

Table 6.4: Suite of Parameter Settings in Study 1.

Setting No.	$\tilde{u}^f$	j	p_j	r_j	h_j	θ^j	c^j
1	4.0	1	5.5	5.0	0.5	0.80	3.0
		2	4.0	6.0	0.7	0.80	3.5
2	4.0	1	5.5	5.0	0.5	0.85	3.0
		2	4.0	6.0	0.7	0.90	3.5
3	4.0	1	5.5	5.0	0.5	0.90	3.0
		2	4.0	6.0	0.7	0.85	3.5
4	4.0	1	5.5	5.0	0.5	0.95	3.0
		2	4.0	6.0	0.7	0.95	3.5
5	4.0	1	5.5	15.0	0.5	0.85	3.0
		2	4.0	20.0	0.7	0.90	3.5
6	4.2	1	5.0	8.0	0.5	0.85	4.0
		2	8.0	10.0	0.7	0.90	4.0
7	4.5	1	7.0	8.0	0.9	0.95	3.8
		2	8.0	10.0	0.7	0.90	3.5
8	5.0	1	7.0	8.0	0.9	0.85	3.8
		2	8.0	10.0	0.7	0.85	3.5

distribution of their unreliable suppliers. Hence, throughout the paper, we concentrate more on the cases where firms have errors in their steady state beliefs. Settings 3-6 include variations on the other parameters as well and, for consistency, are also chosen to be the same as settings 5-8 of Table 6.4.

(ii) Suite of Transition Probability Matrix (t.p.m.) Settings

Let the state space for the Discrete Time Markov Chain (DTMC) dynamics of the threat levels of the unreliable suppliers be based on the S&P credit risk rating system of the firms with state space

$$\{1=\text{AAA}, 2=\text{AA}, 3=\text{A}, 4=\text{BBB}, 5=\text{B/BB}, 6=\text{CCC/CC/C}\} \cup \{0 = Default\}$$

for which Markov chain modeling is commonly used (it is noteworthy that Moody's

Table 6.5: Suite of Parameter Settings in Study 2.

Setting No.	$\tilde{u}^f$	j	p_j	r_j	h_j	θ^j	c^j
1	4.0	1	5.5	5.0	0.5	0.80	3.0
		2	4.0	6.0	0.7	0.80	3.5
2	4.0	1	5.5	5.0	0.5	0.92	3.0
		2	4.0	6.0	0.7	0.92	3.5
3	4.0	1	5.5	15.0	0.5	0.85	3.0
		2	4.0	20.0	0.7	0.90	3.5
4	4.2	1	5.0	8.0	0.5	0.85	4.0
		2	8.0	10.0	0.7	0.90	4.0
5	4.5	1	7.0	8.0	0.9	0.95	3.8
		2	8.0	10.0	0.7	0.90	3.5
6	5.0	1	7.0	8.0	0.9	0.85	3.8
		2	8.0	10.0	0.7	0.85	3.5

bond rating system also has a similar seven state structure). Then let:

$$A = \begin{bmatrix} 0.10 & 0.40 & 0.20 & 0.10 & 0.10 & 0.10 \\ 0.10 & 0.10 & 0.40 & 0.20 & 0.10 & 0.10 \\ 0.15 & 0.05 & 0.10 & 0.30 & 0.20 & 0.20 \\ 0.20 & 0.10 & 0.10 & 0.10 & 0.30 & 0.20 \\ 0.25 & 0.10 & 0.10 & 0.10 & 0.20 & 0.25 \\ 0.30 & 0.10 & 0.10 & 0.10 & 0.20 & 0.20 \end{bmatrix}$$

and

$$B = \begin{bmatrix} 0.03 & 0.37 & 0.30 & 0.15 & 0.10 & 0.05 \\ 0.05 & 0.35 & 0.30 & 0.15 & 0.10 & 0.05 \\ 0.07 & 0.33 & 0.30 & 0.15 & 0.10 & 0.05 \\ 0.10 & 0.30 & 0.10 & 0.30 & 0.10 & 0.10 \\ 0.12 & 0.28 & 0.30 & 0.10 & 0.10 & 0.10 \\ 0.20 & 0.10 & 0.10 & 0.20 & 0.20 & 0.20 \end{bmatrix}$$

be two different transition probability matrices (t.p.m.'s) on this state space. Note that by solving the balance equations for t.p.m. A we have $\pi_0 \simeq 0.185$ but for t.p.m. B we have $\pi_0 \simeq 0.080$. Hence, A represents the dynamics of a relatively low steady state reliability and B represents a higher steady state reliability. Moreover, notice that in both t.p.m.'s A and B, a higher threat level represents a higher chance of disruption in the next period. Also, these t.p.m.'s are chosen to reflect the real-world

dynamics of disruptions of a firm where a firm may stay in the same disruption risk level in the next period, go to a higher level, or to a lower one. Obviously, Bernoulli dynamics are a special case of our general DTMC setting.

Table 6.6: Suite of t.p.m. Settings.

t.p.m. Setting	$\mathbf{W}^1$	$\mathbf{W}^2$	$1-\pi_0^1$	$1-\pi_0^2$
1	A	A	0.815	0.815
2	A	B	0.815	0.920
3	B	A	0.920	0.815
4	B	B	0.920	0.920
5	C	C	0.815	0.815
6	C	D	0.815	0.920
7	D	C	0.920	0.815
8	D	D	0.920	0.920

For our analysis we consider a suite of t.p.m. settings as is illustrated in the following table, where $\mathbf{W}^j$ and $1-\pi_0^j$, respectively denote the t.p.m. and steady state reliability of supplier j. These settings in addition to the different reliability beliefs presented in Tables 6.4 and 6.5 generate a test suite of different belief errors.

BIBLIOGRAPHY

[1] N. Agrawal and S. Nahmias. Rationalization of the supplier base in the presence of yield uncertainty. *Production Operation Management*, 6:291–308, 1997.

[2] O.Z. Aksin and F. Karaesmen. Designing flexibility: Characterizing the value of cross-training practices. working paper, INSEAD, Fontainebleau Cedex, France, 2002.

[3] G. Allon, S. Deo, and W. Lin. The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. Working Paper, Kellogg School of Business., 2010.

[4] American College of Emergency Physicians. American college of emergency physicians national report card on the state of emergency medicine. Available at http://www.acep.org/assets/0/16/648/1994/00FA9DFA-9B89-4DA8-A3D8-5FBD37DD858D.pdf, 2006.

[5] American Hospital Association. Emergency Department overload: a growing crisis. The results of the American Hospital Association survey of Emergency Department (ED) and hospital capacity. Falls Church, VA: American Hospital Association, 2002, 2002.

[6] S. Andradottir, H. Ayhan, and D. G. Down. Compensating for Failures with Flexible Servers. *Oper. Res.*, 55(4):753–768, 2007.

[7] S. Andradottir, H. Ayhan, and D. G. Down. Design principles for fexible systems. Working Paper, 2010.

[8] S. Andradóttir, H. Ayhan, and D.G. Down. Server assignment policies for maximizing the steady-state throughput of finite queueing systems. *Management Sci.*, 47(10):1421–1439, 2001.

[9] S. Andradóttir, H. Ayhan, and D.G. Down. Dynamic server allocation for queueing networks with flexible servers. *Oper. Res.*, 51(6):952–968, 2003.

[10] R. Anupindi and R. Akella. Diversification under supply uncertainty. *Management Sci.*, 39(8):944–963, 1993.

[11] N.T. Argon and S. Ziya. Priority assignment under imperfect information on customer type identities. *Manufacturing Service Oper. Management*, 11(4):674–693, 2009.

[12] M. Armony and N. Bambos. Queueing dynamics and maximal throughput scheduling in switched processing systems. *Queueing Systems: Theory Appl.*, 44(3):209–252, 2003.

[13] V. Babich. Vulnerable options in supply chains: Effects of supplier competition. Naval Research Logistics (forthcoming), 2007.

[14] V. Babich, A. Burnetas, and P. Ritchken. Competition and diversification effects in supply chains with supplier default risk. Manufacturing and Service Operations Management (forthcoming), 2007.

[15] A. Bassamboo, R.S. Randhawa, and J.A. Van Mieghem. A little flexibility is all you need: Asymptotic optimality of tailored chaining and pairing in queuing systems. Working Paper, Kellog School of Business, Northwestern University, 2009.

[16] S.L. Bell and R.J. Williams. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *Annals of Appl. Prob.*, 11(3):608–649, 2001.

[17] S.L. Bell and R.J. Williams. Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: Asymptotic optimality of a threshold policy. *Electronic J. Prob.*, 10:1044–1115, 2005.

[18] D. I. Ben-Tovim, J. E. Bassham, D. M. Bennett, M. L. Dougherty, M. A. Martin, S. J. ONeill, J. L. Sincock, and M. G. Szwarcbord. Redesigning care at the Flinders Medical Centre: clinical process redesign using lean thinking. *Medical Journal of Australia*, 188(6):27–31, 2008.

[19] T. Bielecki and P. R. Kumar. Optimality of zero-inventory policies for unreliable manufacturing systems. *Oper. Res.*, 36:532–541, 1988.

[20] R. Bollapragada, U. S. Rao, and J. Zhang. Managing inventory and supply performance in assembly systems with random supply capacity and demand. *Management Sci.*, 50:1729–1743, 2004.

[21] M. Bramson and R.J. Williams. On dynamic scheduling of stochastic networks in heavy traffic and some new results for the workload process. *Proc. 39th IEEE Conf. Decision and Control, Sydney, Australia*, pages 516–521, 2000.

[22] C. Buyukkoc, P. Varaiya, and J. Walrand. The $c\mu$ rule revisited. *Adv. Appl. Prob.*, 17:237–238, 1985.

[23] S.H. Cho and C.S. Tang. Advance selling in a supply chain under uncertain supply and demand. Working Paper, Carnegie Mellon University.

[24] S.H. Cho and C.S. Tang. Inventory models with random yield. Ph.D. dissertation (Chapter 4).

[25] S. Chopra, G. Reinhardt, and U. Mohan. The importance of decoupling recurrent and disruption risks in a supply chain. *Naval Res. Logist.*, 54:544–555, 2007.

[26] M.C. Chou, G.A. Chua, and Zheng H. Teo, C.-P. Design for process flexibility: Efficiency of the long chain and sparse structure. *Oper. Res.*, 58(1):43–58, 2010.

[27] F.W. Ciarallo, R. Akella, and T. E. Morton. periodic review, production planning model with uncertain capacity and uncertain demand: Optimality of extended myopic policies. *Management Sci.*, 40:320–332, 1994.

[28] A. Cobham. Priority assignment in waiting line problems. *J. Oper. Res. Soc. of Amer.*, 2:70–76, 1954.

[29] A. Cobham. Priority assignment - a correction. *J. Oper. Res. Soc. of Amer.*, 3:547, 1955.

[30] J. Cochran and K.T. Roche. A multi-class queueing network analysis methodology for improving hospital Emergency Department performance. *Comput. and Opers. Res.*, 36(1):1497–1512, 2009.

[31] D.R. Cox and W.L. Smith. *Queues*. Methuen & Co, London, 1961.

[32] M. Dada, N. Petruzzi, and L. Schwarz. A newsvendor's procurement problem when suppliers are unreliable. *Manufacturing Service Oper. Management*, 9:9–32, 2007.

[33] J. G. Dai and W. Lin. Maximum Pressure Policies in Stochastic Processing Networks. *Oper. Res.*, 53(2):197–218, 2005.

[34] J.G. Dai. *Stability of Fluid and Stochastic Processing Networks*. Publication 9, Centre for Mathematical Physics and Stochastics, 1999.

[35] J.G. Dai and S. P. Meyn. Stability and convergence of moments for multiclass queueing networks via fluid models. *IEEE Trans. Automatic Control*, 40:1889–1904, 1995.

[36] D.J. Daley. Stochastically monotone markov chains. *Probability Theory and Related Fields*, 10:305–317, 2009.

[37] C. Dawson. Quake still rattles suppliers. *The Wall Street Journal, Sep. 29*, 2011.

[38] G.A. DeCroix and A. Arreola-Risa. Optimal production and inventory policy for multiple products under resource constraints. *Management Sci.*, 44:950–961, 1998.

[39] D.G. Down and M.E. Lewis. A call center model with upgrades. Working Paper, Department of Computing and Software, McMaster University, 2009.

[40] I. Duenyas, D. Gupta, and T. Olsen. Control of a single-server tandem queueing system with setups. *Oper. Res.*, 46:218–230, 1998.

[41] G.D. Eppen and A.V. Iyer. Lot sizing with random yield: A review. *Oper. Res.*, 43:311–334, 1995.

[42] G.D. Eppen and A.V. Iyer. Backup agreements in fashion buying: the value of upstream fexibility. *Management Sci.*, 43:1469–1484, 1997.

[43] A. Federgruen and N. Yang. Optimal supply diversification under general supply risks. *Oper. Res.*, 57:1451–1468, 2007.

[44] C.M. Fernandes, M.R. Daya, S. Barry, and N. Palmer. Emergency Department patients who leave without seeing a physician: The toronto hospital experience. *Annals of Emrgency Medicine*, 24(6):1092–1096, 1994.

[45] C.M. Fernandes, P. Tanabe, and N. Gilboy et al. Five level triage: a report from the ACEP/ ENA five-level task force. *J. Emerg. Nurs.*, 31:39–50, 2005.

[46] C.H. Fine and R.M. Freund. Optimal investment in product flexible manufacturing capacity. *Management Sci.*, 36:449–466, 1990.

[47] G. FitzGerald, G.A. Jelinek, D. Scott, and M.F. Gerdtz. Emergency Department triage revisited. *Emergency Medicine Journal*, 27:86–92, 2010.

[48] Y. Gerchak and M. Parlar. Yield randomness, cost trade-offs and diversification in the eoq model. *Naval Research Logistics*, 37:341–354, 1990.

[49] D. Gerwin. Manufacturing flexibility: A strategic perspective. *Management Sci.*, 39:395–410, 1993.

[50] N. Gilboy, P. Tanabe, D.A. Travers, A.M. Rosenau, and D.R. Eitel. *Emergency Severity Index, Version 4: Implementation Handbook*. Agency for Healthcare Research and Quality Publication No. 05-0046-2, Rockville, MD, 2005.

[51] J.A. Gordon, J. Billings, and B.R. Asplin et al. Safety net research in emergency medicine: proceedings of the academic emergency medicine consensus conference on the unraveling safety net. *Acad. Emerg. Med.*, 8:10249, 2001.

[52] L. G. Graff, S. Wolf, R. Dinwoodie, D. Buono, and D. Mucci. Emergency physician workload: A time study. *Annals of Emergency Medicine*, 22(7):1156–1163, 1993.

[53] S.C. Graves and B.T. Tomlin. Process flexibility in supply chains. *Management Sci.*, 49:907–919, 2003.

[54] L. V. Green, J. Soares, J. F. Giglio, and R.A. Green. Using queuing theory to increase the effectiveness of Emergency Department provider staffing. *Academic Emergency Medicine*, 13(1):61–68, 2006.

[55] D. Greising and J. Johnsson. Behind boeing 787 delays: problems at one of the smallest suppliers in dreamliner program causing ripple effect. *Chicago Tribune*, (Dec. 08), 2007.

[56] X. Guo and K. Liu. A note on optimality conditions for continuous-time Markov Decision Processes with aevrage cost criterion. *IEEE Trans. on Aut. Contr.*, 46(12):1984–1989, 2001.

[57] D. Gupta. The (Q, r) inventory system with an unreliable supplier. *INFOR*, 34:59–76, 1996.

[58] U. Gurler and M. Parlar. An inventory problem with two randomly available suppliers. *Oper. Res.*, 45:904–918, 1997.

[59] H. Gurnani, R. Akella, and J. Lehoczky. Supply management in assembly systems with random yield and random demand. *IIE Trans.*, 32:701–714, 2000.

[60] J.H. Han, D.J. France, and S.R. Levin et. al. The effect of physician triage on Emergency Department length of stay. *J. of Emerg. Med.*, 39(2):227–233, 2010.

[61] J.M. Harrison. Heavy traffic analysis of a system with parallel servers: Asymptotic analysis of discrete-review policies. *Annals of Appl. Prob.*, 8(3):822–848, 1998.

[62] J.M. Harrison and M.J. López. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems*, 33:339–368, 1999.

[63] E. Hay, L. Bekerman, G. Rosenberg, and R. Peled. Quality assurance of nurse triage: Consistency of results over three years. *American J. of Emrg. Med.*, 19(2):113–117, 2001.

[64] K.B. Hendricks and V.R. Singhal. An empirical analysis of the effect of supply chain disruption on long-run stock price performance and equity risk of the firm. *Prod. and Oper. Mang.*, 14(1):35–52, 2005.

[65] M. Henig, Y. Gerchak, R. Ernst, and D.F. Pyke. An inventory model embedded in designing a supply contract. *Management Sci.*, 43:184–189, 1997.

[66] A. Holdgate, J. Morris, M. Fry, and M. Zecevic1. Accuracy of triage nurses in predicting patient disposition. *Emergency Medicine Australasia*, 19:341–345, 2007.

[67] L.L. Holland, L.L. Smith, and K.E. Blick. Reducing laboratory turnaround time outliers can reduce Emergency Department patient length of stay. *Amer. J. Clin. Path.*, 124:672–674, 2005.

[68] J.C. Hollingsworth, C.D. Chisholm, B.K. Giles, W.H. Cordell, and D.R. Nelson. How do physicians and nurses spend their time in the Emergency Department. *Annals of Emergency Medicine*, 31(1):87–91, 1998.

[69] N. R. Hoot and D. Aronsky. Systematic review of Emergency Department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine*, 52(2):126–136, 2008.

[70] W.J. Hopp, S.M.R. Iravani, and B. Shou. Serial agile production systems with automation. *Oper. Res.*, 53(5):852–866, 2005.

[71] W.J. Hopp, E. Tekin, and M.P. Van Oyen. Benefits of skill chaining in production lines with cross-trained workers. *Management Sci.*, 50(1):83–98, 2004.

[72] W.J. Hopp and M.P. Van Oyen. Agile workforce evaluation: A framework for cross-training and coordination. *IIE Transactions*, 36(10):919–940, 2004.

[73] W.J. Hopp and Z. Yin. Protecting supply chain networks against catasrophic failures. Working Paper, IEMS Dept., Northwestern University, 2011.

[74] E.E. Howell, E.D.Bessman, and H.R. Rubin. Hospitals and innovative Emergency Department admission process. *Journal of General Internal Medicin*, 19:266–2686, 2004.

[75] B. Hu and S. Benjafar. Partitioning of servers in queueing systems during rush hour. *Manufacturing Service Oper. Management*, 11(3):416–428, 2009.

[76] K.V. Ierson and J.C. Moskop. Triage in medicine, part I: Concept, history, and types. *Annals of Emrg. Med.*, 49(3):275–281, 2007.

[77] Institute of Medicine. *Hospital-Based Emergency Care: At the Breaking Point.* National Academies Press, London, 2007.

[78] S.M. Iravani, B. Kolfal, and M.P. Van Oyen. Capability flexibility: A decision support methodology for production and service systems with flexible resources. Working paper. Department of IE/MS, Northwestern University, 2006.

[79] S.M. Iravani, B. Kolfal, and M.P. Van Oyen. Capability flexibility: A decision support methodology for production and service systems with exible resources. *IIE Trans.*, 43:363–382, 2011.

[80] S.M.R Iravani, B. Kolfal, and M.P. Van Oyen. Call center labor cross-training: It's a small world after all. *Management Sci.*, 53(7):1102–1112, 2007.

[81] S.M.R. Iravani, M.P. Van Oyen, and K.T. Sims. Structural flexibility: A new perspective on the design of manufacturing and service operations. *Management Sci.*, 51(2):151–166, 2005.

[82] S.M.R. Iravani, M.P. Van Oyen, and K.T. Sims. Structural flexibility: A new perspective on the design of manufacturing and service operations. *Management Sci.*, 51(2):151–166, 2005.

[83] N.K. Jaiswal. *Priority Queues*. Academic Press, New York, New York, 1968.

[84] W.J. Jordan and S.C. Graves. Principles on the benefits of manufacturing process flexibility. *Management Sci.*, 41(4):577–594, 1995.

[85] W.J. Jordan and S.C. Graves. Principles on the benefits of manufacturing process flexibility. *Management Sci.*, 41(4):577–594, 1995.

[86] J.S. Kakalik and J.D.C. Little. *Optimal Service Policy for the M/G/1 Queue with Multiple Classes of Arrival*. Rand Corporation Report, 1971.

[87] W.W. Keen. *The Treatment of War Wounds*. W.B. Saunders, Philadelphia, PA, 1917.

[88] R.K. Khare, E.S. Powell, G. Reinhardt, and M. Lucenti. Adding more beds to the Emergency Department or reducing admitted patient boarding times: Which has a more significant influence on Emergency Department congestion? *Annals of Emergency Medicine*, 53(5):575–585, 2008.

[89] C.R. Kim. Toyota aims for quake-proof supply chain. *Reuters, Sep. 06*, 2011.

[90] D. L. King, D. I. Ben-Tovim, and J. Bassham. Redesigning Emergency Department patient flows: Application of lean thinking to health care. *Emergency Medicine Australasia*, 18:391–397, 2006.

[91] L. Kinsman, R. Champion, G. Lee, M. Martin, K. Masman, E. May, T. Mills, M.D. Taylor, P. Thomas, R.J. Williams, and S. Zalstein. Assessing the impact of streaming in a regional Emergency Department. *Emergency Medicine Australasia*, 20:221–227, 2008.

[92] P.R. Kleindorfer and G.H. Saad. Managing disruption risk in supply chain. *Prod. and Oper. Mang.*, 14(1):53–58, 2005.

[93] P. Kouvelis and G. Vairaktarakis. Flowshops with processing flexibility across production stages. *IIE Trans.*, 30:735–746, 1998.

[94] S.L. Kronick and J.S. Desmond. Blink: Accuracy of physician estimates of patient disposition at the time of ED triage. SAEM Midwest Regional Meeting, 2009.

[95] H.L. Lee, V. Padmanabhan, and S. Whang. Information distortion in a supply chain: The bullwhip effect. *Management Sci.*, 43:546–558, 1997.

[96] P. A.W. Lewis and G. S. Shedler. Simulation of nonhomogenous Poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, 1979.

[97] P. A.W. Lewis and G. S. Shedler. Simulation of nonhomogenous Poisson processes with degree-two exponential polynomial rate function. *Oper. Res.*, 27(5):1026–1039, 1979.

[98] D. Liew, D. Liew, and M.P. Kennedy. Emergency Department length of stay independently predicts excess inpatient length of stay. *Medical Journal of Australia*, 179(17):524–526, 2003.

[99] S. Lippman. Applying a new device in the optimization of exponential queueing system. *Oper. Res.*, 23(4):687–710, 1975.

[100] S.W. Liu, S.H. Thomas, J.A. Gordon, and J. Weissman. Frequency of adverse events and errors among patients boarding in the emergency department. *Acad. Emerg. Med.*, 12:49b–50b, 2005.

[101] A. Mandelbaum and M.I. Reiman. On pooling in queueing networks. *Management Sci.*, 44(7):971–981, 1997.

[102] A. Mandelbaum and A.L. Stolyar. Scheduling flexible servers with convex delay costs: Heavy-taffic optimality of the generalaized $c\mu$-rule. *Oper. Res.*, 52(6):836–855, 2004.

[103] L.F. McCaig and N. Ly. National hospital ambulatory medical care survey: 2000 Emergency Department summary. *National Health Statistics Report*, pages 1–31, 2002.

[104] R.R. Meyer, Rothkopf M. H., and S. A. Smith. Reliability and inventory in a productionstorage system. *Management Sci.*, 25:799–807, 1979.

[105] Sean P. Meyn. Sequencing and routing in multiclass queueing networks part ii: Workload relaxations. *SIAM Journal on Control and Optimization*, 42(1):178–217, 2003.

[106] O. Miro, M. Sanchez, G. Espinosa, B Coll-Veient, E. Bragulat, and J. Milla. Analysis of patient flow in the Emergency Department and the effect of an extensive reorganisation. *Emergency Medicine Journal*, 20:143–148, 2003.

[107] K. Moinzadeh and P. Aggrawal. Analysis of a production/ inventory system subject to random disruptions. *Management Sci.*, 43:1577–1588, 1997.

[108] K. Moinzadeh and P. Aggrawal. Inventory management under random disruptions and partial back-orders. *Naval Res. Logist.*, 45:687–703, 1998.

[109] J.C. Moskop and K.V. Ierson. Triage in medicine, part II: Underlying values and principles. *Annals of Emrg. Med.*, 49(3):282–287, 2007.

[110] M. Parlar. Continuous review inventory problem with random supply interruptions. *Eur. J. Oper. Res.*, 99:366–385, 1997.

[111] M. Parlar and D. Berkin. Future supply uncertainty in eoq models. *Naval Res. Logist.*, 38:107–121, 1991.

[112] M. Parlar and D. Perry. Analysis of a (q, r, t) inventory policy with deterministic and random yields when future demand is uncertain. *Eur. J. Oper. Res.*, 84:431–443, 1995.

[113] M. Parlar and D. Perry. Inventory models of future supply uncertainty with single and multiple suppliers. *Naval Research Logistics*, 43:191–210, 1996.

[114] S. R. Pitts, R. W. Niska, J. Xu, and C. W. Burt. National hospital ambulatory medical care survey: 2006 Emergency Department summary. *National Health Statistics Report*, 7:1–39, 2008.

[115] V. Reitman. Behind boeing 787 delays: problems at one of the smallest suppliers in dreamliner program causing ripple effect. *Wall Street Journal*, (May. 08), 1997.

[116] D. Richardson. Reducing patient times in Emergency Department. *Medical Journal of Australia*, 179(17):516–517, 2003.

[117] D. Richardson. Increase in patient mortality at 10 days associated with Emergency Department overcrowding. *Medical Journal of Australia*, 184:213–216, 2006.

[118] M.H. Rothkopf and P. Rech. Perspective on queueing: Combining queues is not always beneficial. *Oper. Res.*, 35:906–909, 1987.

[119] S. Russ, I. Jones, and D. Aronsky et. al. Placing physician orders at triage: The effect on length of stay. *Annals of Emerg. Med.*, 56(1):27–33, 2010.

[120] S. Saghafian, W.J. Hopp, M.P. Van Oyen, J.S. Desmond, and S.L. Kronick. Complexity-based triage: A tool for improving patient safety and operational efficiency. Working Paper, Dept. of Industrial and Operations Eng., University of Michigan, 2012.

[121] S. Saghafian, W.J. Hopp, M.P. Van Oyen, J.S. Desmond, and S.L. Kronick. Patient streaming as a mechanism for improving responsiveness in Emergency Departments. *Oper. Res. (forthcoming)*, 2012.

[122] S. Saghafian and M.P. Van Oyen. The value of flexible backup suppliers and disruption risk information: Newsvendor analyses with recourse. *IIE Transactions (forthcoming)*.

[123] S. Saghafian and M.P. Van Oyen. Compensating for dynamic supply disruptions with backup flexibility. Working paper, Univ. of Michigan, Dept. of Indust. and Opers. Eng., 2012.

[124] S. Saghafian, M.P. Van Oyen, and B. Kolfal. The "W" network and the dynamic control of unreliable flexible servers. *IIE Transactions*, 43(12):893–907, 2011.

[125] M.J. Schull, A. Kiss, and J.-P. Szali. The effect of low complexity patients on Emergency Department waiting times. *Annals of Emergency Medicine*, 49(3):257–264, 2007.

[126] L.I. Sennott. *Stochastic Dynamic Programming and the Control of Queueing Systems*. Wiley Series in Probability and Statistics, John Wiley and Sons, New York, 1999.

[127] D.A. Serel, M. Dada, and H. Moskowitz. Sourcing decision with capacity reservation contract. *Eur. J. Opers. Res.*, 131:635–648, 2001.

[128] A.K. Sethi and S.P. Sethi. Flexibility in manufacturing: A survey. *The International Journal of Flexible Manufacturing Systems*, 2:289–328, 1990.

[129] Y. Sheffi. *The resilient enterprise: overcoming vulnerability for competitive advantage*. MIT Press, Cambridge, Massachusetts, 2007.

[130] K. Siddharathan, W.J. Jones, and J.A. Johnson. A priority queueing model to reduce waiting times in emergency care. *International J. of Health Care Quality Assurance*, 9(5):10–16, 1996.

[131] R. Smallwood and E. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Oper. Res.*, 21:1071–1088, 1973.

[132] L. I. Solberg, B. R. Asplin, and D. J. Magid R. M. Weinick. Emergency Department crowding: Consensus development of potential measures. *Annals of Emergency Medicine*, 42(6):824–834, 2003.

[133] J. Song and P. Zipkin. Inventory control with information about supply conditions. *Management Sci.*, 42(10):1409–1419, 1996.

[134] R. SoRelle. Homicide charges against Illinois ED stun EM. *Emergency Medicine News*, 28(12):1,25, 2006.

[135] R. SoRelle. Breaking news: Health reform and the ED: Prepare for the surge. *Emergency Medicine News*, 32(5):1,20, 2010.

[136] P.C. Sprivulis, J.A. Da Silva, and I.G Jacobs et al. The association between hospital overcrowding and mortality among patients admitted via Western Australian Emergency Departments. *Medical Journal of Australia*, 184:208–212, 2006.

[137] M. S. Squillante, C. H. Xia, D. D. Yao, and L. Zhang. Threshold-based priority policies for parallel-server systems with affinity scheduling. *Proc. 2001 Amer. Control Conf., Arlington, VA*, pages 2992–2999, 2001.

[138] S.J. Steindel and P.J. Howanitz. Changes in Emergency Department turnaround time performance from 1990 to 1993. *Arch. Path. & Lab. Med.*, 121:1031–1041, 1997.

[139] F.F. Suarez, M.A. Cusumano, and C.H. Fine. An empirical study of flexibility in manufacturing. *Sloan Management Rev.*, 37:25–32, 1995.

[140] J.M. Swaminathan and J.G. Shanthikumar. Supplier diversification: effect of discrete demand. *Oper. Res. Let.*, 25:213–221, 1999.

[141] R. Swinney and S. Netessine. Long-term contracts under the threat of supplier default. *Manufacturing Service Oper. Management*, 11:109–127, 2009.

[142] E. Tekin, W.J. Hopp, , and M.P. Van Oyen. Pooling strategies for service center agent cross-training. *IIE Transactions*, 41:546–561, 2009.

[143] E.J. Thomas, D.M. Studdert, and H.R. Burstin et. al. Incidene and type of adverse events and negligent care in Utah and Colorado. *Medical Care*, 38(3):261–271, 2000.

[144] B. Tomlin. The impact of supply-learning on a firms sourcing strategy and inventory investment when suppliers are unreliable. *Manufacturing Service Oper. Management*, 11:192–209, 2009.

[145] B. Tomlin and Lawrence Snyder. On the value of a threat advisory system for managing supply chain disruptions. Working Paper, 2006.

[146] B. Tomlin and Lawrence Snyder. On the value of a threat advisory system for managing supply chain disruptions. Working Paper, 2006.

[147] B. Tomlin and Y. Wang. On the value of mix flexibility and dual sourcing in unreliable newsvendor networks. *Manufacturing Service Oper. Management*, 7(1):37–57, 2005.

[148] S. Trzeciak and E.P. Rivers. Emergency Department overcrowding in the United States: an emerging threat to patient safety and public health. *Emerg. Med. J.*, 20(5):402–405, 2003.

[149] S.P. van der Zee and H. Theil. Priority assignment in waiting-line problems under conditions of misclassification. *Oper. Res.*, 9:875–885, 1961.

[150] N.M. Van Dijk and E. Van Der Sluis. To pool or not to pool in call centers. *Prod. Oper. Man.*, 17(3):296–305, 2008.

[151] J. A. Van Mieghem. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Annals of Appl. Prob.*, 5(3):809–833, 1995.

[152] M.P. Van Oyen, E.G.S. Gel, and W. J. Hopp. Opportunity for workforce agility in collaborative and non-collaborative work systems. *IIE Transactions*, 33(9):761–777, 2001.

[153] J. Vance and P. Spirvulis. Triage nurses validly and reliably estimate Emergency Department patient complexity. *Emergency Medicine Australasia*, 17:382–386, 2005.

[154] J. A. VanMieghem. Investment strategies for flexible resources. *Management Sci.*, 44(8):1071–1078, 1998.

[155] Michael H. Veatch. A $c\mu$ rule for prallel servers with two tiered $c\mu$ preference. Working Paper, Math. Dept., Gordon College, 2010.

[156] J. Walrand. *An Introduction to Queueing Networks*. Prentice-Hall, New York, 1988.

[157] Q. Wang. Modeling and analysis of high risk patient queues. *Eur. J. of Oper. Res.*, 155:502–515, 2004.

[158] Y. Wang, W. Gilland, and B. Tomlin. Mitigating supply risk: dual sourcing or process improvement? *Manufacturing Service Oper. Management*, 12:489–510, 2010.

[159] S. J. Welch. Patient segmentation: Redesigning flow. *Emergency Medicine News*, 31(8), 2008.

[160] S.J. Welch and S.J. Davidson. The performance limits of traditional triage. *Annals of Emerg. Med.*, 58(2):143–144, 2011.

[161] W. Whitt. Partitioning customers into service groups. *Management Sci.*, 45:1579–1592, 1999.

[162] R.W. Wolf. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Englewood Cliffs, New Jersey, 1989.

[163] Z. Yang, G. Aydin, V. Babich, and D. Beil. Supply disruptions, asymmetric information, and a backup production option. *Management Sci.*, 55:192–209, 2009.