

Stochastic Modeling and Analysis of Cooperative Production and Service Networks

by

Behnaz Hosseini

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy

in

Industrial Engineering and Operations Management



**KOÇ
UNIVERSITY**

January 2019

**Stochastic Modeling and Analysis of Cooperative Production and Service
Networks**

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Behnaz Hosseini

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

Prof. Dr. Barış Tan

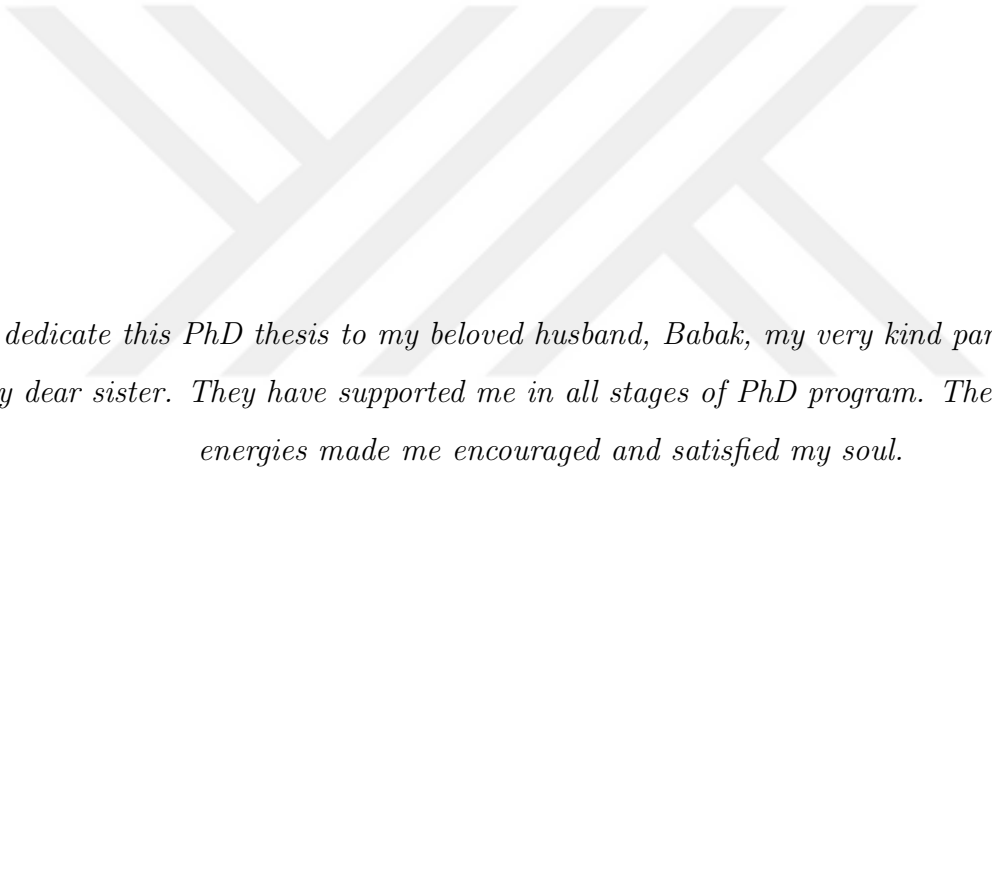
Prof. Dr. Fikri Karaesmen

Prof. Dr. Sencer Yeralan

Prof. Dr. Lerzan Örmeci

Assoc. Prof. Dr. Aybek Korugan

Date: _____



I dedicate this PhD thesis to my beloved husband, Babak, my very kind parents, and my dear sister. They have supported me in all stages of PhD program. Their positive energies made me encouraged and satisfied my soul.

ABSTRACT

In response to the economic condition and competition in the business environment, manufacturers and service providers began using cooperation-based business models to utilize their capacity effectively and access more customers. Different mechanisms enable the producers or the service providers to work together and share resources. These mechanisms facilitate cooperation among different members and provide more efficient services at a lower cost. This thesis consists of three parts that address modeling and analysis of cooperative production and service networks.

In the first part, we study the cooperative production business model for a group of producers participating in a network to receive additional external customers who buy products under certain conditions. The cooperative network routes an arriving external customer to one of the participants based on the network routing policy. We model the system as a Markov decision process and determine the optimal production and rationing policies for each participating manufacturer as well as the optimal routing policy for the network by using mathematical programming approach. Our results show that a cooperative network can provide financial benefits for the producers and also allows the external customers to obtain goods at a lower price with the desired service level.

In the second part, the mathematical programming approach is proposed as a viable alternative method for simulation, performance evaluation, and optimization of a general continuous-flow production system with an intermediate finite buffer. We show that a simulated sample realization of the system is obtained by solving a mixed-integer linear program, and our numerical results present a significant computational improvement compared to using a discrete-event simulation.

In the last part of the dissertation, we examine the operational model of a coop-

erative service network with an application in telemedicine platform. The platform consists of a number of doctors serving their own patients in person, and receive the on-line medical requests from the platform whenever they are available. We consider the dynamic admission problem for each doctor in the platform who faces with heterogeneous patient streams. This problem is modeled as a Markov decision process, and we prove the existence of threshold-type optimal control policy. Once the optimal operation of the doctors is modeled, we then analyze the operation of the whole platform. This analysis allows us to design a platform effectively by choosing the right number of doctors based on the doctors' utilization, the patients' arrival rate and the desired service level. We show that a cooperative service network such as a telemedicine platform is an effective way of utilizing the idle capacity of the servers and serving the patients with the desired level.

We believe that the results derived in this dissertation will be of value for the producers and service providers planning to join the cooperative networks to reach the full benefits of cooperation. This dissertation also extends the state of art in the analyses of cooperative networks by considering different kinds of settings that have not been addressed in the literature. Therefore, we provide researchers a basis for further research and development on the new cooperation based business models both for the production and service systems.

ÖZETÇE

İş ortamındaki ekonomik duruma ve rekabete cevap olarak, üreticiler ve servis sağlayıcılar kapasitelerini etkin bir şekilde kullanmak ve daha fazla müşteriye erişmek için işbirliğine dayalı iş modelleri kullanmaya başlamışlardır. Farklı mekanizmalar üreticilerin veya servis sağlayıcıların birlikte çalışmasını ve kaynakları paylaşmasını sağlar. Bu mekanizmalar farklı üyeler arasında işbirliğini kolaylaştırır ve düşük maliyetle daha verimli hizmetler sunar. Bu tez, kooperatif üretim ve servis ağlarının modellenmesi ve analizini ele alan üç bölümden oluşmaktadır.

İlk bölümde, belirli koşullar altında ürün satın alan ek dış müşterileri de çekebilmek amacıyla bir kooperatif üretim ağ iş modelini inceledik. Kooperatif ağı, gelen bir dış müşteriyi, ağ yönlendirme politikasına dayalı olarak katılımcılardan birine yönlendirir. Biz bu sistemi bir Markov karar süreci olarak modelledik ve katılan her üretici için en iyi üretim ve satış politikalarını ve ayrıca matematiksel programlama yaklaşımını kullanarak ağ için en uygun yönlendirme politikasını belirledik. Sonuçlarımız kooperatif ağının üreticiler için finansal faydalar sağlayabildiğini ve dış müşterilerin de istenen hizmet seviyesinde daha düşük bir fiyatla mal edinebileceğini gösterdi.

İkinci bölümde, matematiksel programlama yaklaşımı, genel bir sürekli akış üretim sisteminin simülasyon, performans değerlendirme ve optimizasyonu için uygun bir alternatif yöntem olarak önerilmiştir. Bu bölümde sistemin simülasyonunun karma tam sayılı bir doğrusal program çözümü ile elde edildiğini gösterdik. Sonuçlarımıza göre, bu yöntem kesikli olay simülasyonu kullanmaya kıyasla anlamlı bir hesaplama iyileştirmesi sunmaktadır.

Tezin son bölümünde, teletıp platformunda uygulaması olan bir kooperatif hizmet ağının işletim modelini inceledik. Ele aldığımız teletıp platform modeli, kendi hastalarına hizmet eden ve uygun oldukları zaman platform hastalarını kabul eden birkaç

doktordan oluşmaktadır. Bu çalışmada, platformdaki heterojen hasta akışlarıyla karşı karşıya olan her doktor için dinamik kabul problemini bir Markov karar süreci olarak modelledik ve eşik tipi optimal kontrol politikasının varlığını kanıtladık. Doktorların optimal çalışması modellendikten sonra, tüm platformun çalışması analiz edildi. Bu analiz, doktorların kullanım seviyesine, hastaların varış oranlarına ve istenen hizmet seviyelerine göre doğru sayıda doktor seçerek etkili bir platform tasarlamamızı sağladı. Sonuçlarımız, teletıp platformu gibi bir işbirliği servis ağının, servis sağlayıcıları için boşta kalan kapasitelerini kullanmanın ve hastalara istenen düzeyde hizmet etmenin etkili bir yolu olduğunu göstermektedir.

Bu tezde elde edilen sonuçların işbirliğinin tam yararlarına ulaşmak için kooperatif ağlarına katılmayı planlayan üreticiler ve servis sağlayıcılar için önemli bulgular içerdiğini düşünürüz. Bu tez aynı zamanda, literatürde ele alınmamış farklı türden modelleri dikkate alarak kooperatif ağların analiz tekniklerini genişletmektedir. Bu nedenle, araştırmacılara hem üretim hem de hizmet sistemleri için yeni işbirliğine dayalı iş modelleri hakkında daha fazla araştırma ve geliştirme yapabilmeleri için bir temel sağlamaktayız.

ACKNOWLEDGMENTS

I would like to state my special appreciation and thanks to my supervisor Prof. Barış Tan for the continuous support of my PhD study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. He behaves me such that I always feel free to share with him every issue; not only academic issues but also private life issues. He has set an example of excellence as an advisor, researcher and a role model. Besides my advisor, I would like also express my special thanks to Prof. Fikri Karaesmen, Prof. Lerzan Örmeci, and Prof. Sencer Yeralan as three members of my thesis committee. During last four years, they helped me in different issues and let me meet them for resolving different affairs. Their comments and suggestions were always helpful and informative for providing this thesis. I also appreciate Prof. Aybek Korugan for his participation in the defense meeting and constructive criticisms before and after that.

I would especially like to thank my amazing family for their infinite love, support, and encouragement I have gotten over the years. I undoubtedly could not have done this without them.

Finally, I thank Dr. Oktay Karabağ for his great friendship and help during the entire time of my PhD studies as a wonderful office-mate.

TABLE OF CONTENTS

List of Tables	xiii
List of Figures	xiv
Nomenclature	xix
Chapter 1: Introduction	1
Chapter 2: Modeling and Analysis of a Cooperative Production Network	9
2.1 Past Work	11
2.1.1 Production Control with Multiple Demand Classes	11
2.1.2 Cooperation on Production, Inventory, and Sales	13
2.2 Model Description	14
2.3 Independent Operation (No cooperation case)	17
2.4 Cooperative Network with N Producers and Static Routing Policy . . .	21
2.5 Cooperative Network with N Producers and Dynamic Routing Policy	23
2.5.1 Exact Analysis of the Optimal Production, Rationing and Routing Policies of the Cooperative Network with N Participants Using Linear Programming Formulation	24
2.5.2 Structure of the Optimal Production, Rationing and Routing Control Policies for the Network with Two Participants	27
2.6 Coalition Structure among N Manufacturers	29
2.7 Approximate Analysis of a Dynamic Network with a High Number of Homogeneous Producers	30

2.7.1	Single Queue Approximation	31
2.7.2	Approximate Analysis of the Cooperative Network	33
2.7.3	Accuracy of the Approximation Method	34
2.7.4	Operation of the Cooperative Production Network for a Large Number of Homogeneous Producers	35
2.8	Benefit of Joining a Cooperative Network	36
2.8.1	Comparing the Benefits of the Cooperative Networks with the Dynamic vs. Static Routing Policies and Homogeneous Members	36
2.8.2	Comparing the Benefits of Cooperative Network with the Dy- namic vs. Static Routing Policies and Heterogeneous Members	39
2.8.3	Coalition Structure among Heterogeneous Manufacturers	41
2.8.4	Benefits of a Cooperative Network with a High Number of Homogeneous Participants	43
2.8.5	Concluding Remarks	48
Chapter 3:	Simulation and Optimization of Continuous Flow Pro- duction Systems by Using Mathematical Programming	50
3.1	Past work	51
3.2	General Model	53
3.3	System Dynamics	54
3.3.1	Inputs and State Variables	55
3.3.2	Critical Times	56
3.3.3	Dynamics of Reaching Full Buffer	57
3.3.4	Dynamics of Reaching Empty Buffer	58
3.3.5	Dynamics of Staying at the Same Level	59
3.3.6	Buffer Level Dynamics	59
3.3.7	Example: A Two Station Model with Quality Modeling	60
3.4	Optimization Based Simulation	61

3.4.1	Linear Expressions for the State Variables	62
3.4.2	Performance Measures	64
3.4.3	Structure of the Mathematical Program for Simulation	66
3.4.4	The Structure of Mathematical Programming Problem	68
3.5	Joint Simulation and Optimization	70
3.5.1	Minimum Buffer Capacity Determination Problem to Achieve the Desired Production Rate	70
3.5.2	Optimal Buffer Capacity Determination Problem to Maximize the Average Profit	72
3.6	Numerical Results	73
3.6.1	Validation and Effects of System Parameters	73
3.6.2	Results for a System with Non-exponential Transition Times .	74
3.6.3	Results for Buffer Determination Problems	75
3.6.4	Computational Efficiency of Mathematical Programming Approach	77
Chapter 4:	Modeling and Analysis of a Cooperative Service Net- work	83
4.1	Past Work	86
4.2	Model Description	88
4.2.1	Decision Problems	90
4.3	Analysis of the Doctors' Admission Policy for Platform Patients	91
4.3.1	Structure of the Optimal Scheduling Policy	92
4.3.2	Stationary Analysis of the System	93
4.4	Analysis of the Telemedicine Platform under the Optimal Policy . . .	95
4.5	Numerical Experiments	98
Chapter 5:	Conclusion	103
Chapter 6:	Appendices	109

6.1	Appendix A	109
6.1.1	Performance Measures of the Independent Operation when the External Business Offer Is Accepted	109
6.1.2	Performance Measures of the Independent Operation when the External Business Offer Is Rejected	110
6.1.3	Proof of Proposition 2.1	110
6.1.4	Proof of Proposition 2.2	112
6.2	Appendix B	113
6.2.1	Proof of Proposition 3.1	113
6.2.2	Proof of Proposition 3.2	115
6.3	Appendix c	117
6.3.1	Proof of Theorem 4.1	117
6.3.2	Proof of Proposition 4.1	121

LIST OF TABLES

2.1	The optimal total profit of each alternative group C_k among three heterogeneous producers ($\lambda_1 = 2, \lambda_2 = 3, \lambda_3 = 4, \lambda_{e1} = \lambda_{e2} = 4, \lambda_{e3} = 5, \mu_1 = \mu_2 = 4, \mu_3 = 6, p_1 = 100, p_2 = 90, p_3 = 85, p_{e1} = p_{e2} = 75, p_{e3} = 70, h_1 = h_2 = h_3 = 15, \gamma = 0.6$).	42
2.2	The optimal total profit of possible coalitions among three heterogeneous producers ($\lambda_1 = 2, \lambda_2 = 3, \lambda_3 = 4, \lambda_{e1} = \lambda_{e2} = 4, \lambda_{e3} = 5, \mu_1 = \mu_2 = 4, \mu_3 = 6, p_1 = 100, p_2 = 90, p_3 = 85, p_{e1} = p_{e2} = 75, p_{e3} = 70, h_1 = h_2 = h_3 = 15, \gamma = 0.6$).	43
3.1	Problem Structure Based on the Number of Elements in T^+ , T^- , and T^0	68
3.2	Structure of the optimization problem based on run duration time . .	79
3.3	Comparison of simulation times based on run duration time	82

LIST OF FIGURES

2.1	Accepting external customers directly	16
2.2	Cooperative network for external sales	16
2.3	(a) The boundary between the external customers' accept and reject regions for different values of selling price p_e and requested service level γ when $\rho = 0.4$, (b) The boundary between the external customers' accept and reject regions for different values of selling price p_e and requested service level γ when $\rho = 0.8$, ($\mu = 5, p = 100, h = 5$).	21
2.4	(a) Optimal production, rationing, and routing policy for a network with two homogeneous manufacturers participating ($\lambda_1 = 2, \lambda_2 = 2, \mu_1 = \mu_2 = 4, \lambda_e = 5, p_1 = p_2 = 100, p_{e1} = p_{e2} = 40, h_1 = h_2 = 10, N = 2, \gamma = 0.6$) (b) Optimal production, rationing, and routing policy for a network with two heterogeneous manufacturers participating ($\lambda_1 = \lambda_2 = 2, \mu_1 = \mu_2 = 4, \lambda_e = 5, p_1 = 70, p_2 = 100, p_{e1} = p_{e2} = 35, h_1 = h_2 = 10, N = 2, \gamma = 0.6$).	27
2.5	The percentage error of profit between the exact solution of dynamic network and its approximation using single queue method (Group 1: $\lambda = 3, \lambda_{ei} = 4, p = 100, p_e = 50, h = 15, \mu = 5, \gamma = 0.65$, Group 2: $\lambda = 4, \lambda_{ei} = 6, p = 100, p_e = 50, h = 15, \mu = 7, \gamma = 0.75$, Group 3: $\lambda = 5, \lambda_{ei} = 4, p = 100, p_e = 70, h = 15, \mu = 6, \gamma = 0.8$)	35

2.6	The boundary between the external customers' accept and reject regions for different values of selling price p_e and requested service level γ in independent operation, static and dynamic networks with two identical members ($\lambda_1 = \lambda_2 = 3, \mu_1 = \mu_2 = 5, \lambda_{e1} = \lambda_{e2} = 4, p_1 = p_2 = 100, h_1 = h_2 = 5$).	37
2.7	Profit rate change and its percentage of increase ($\Delta\%$) in response to different values of selling price p_e and specified (γ) in static and dynamic networks with two identical members ($\lambda_1 = \lambda_2 = 3, \mu_1 = \mu_2 = 5, \lambda_{e1} = \lambda_{e2} = 4, p_1 = p_2 = 100, h_1 = h_2 = 5$)	38
2.8	Percentage of profit increase ($\Delta\%$) in response to different values of desired service level γ and selling price p_e in a dynamic network with two identical members ($\lambda_1 = \lambda_2 = 3, \mu_1 = \mu_2 = 5, \lambda_{e1} = \lambda_{e2} = 4, p_1 = p_2 = 100, h_1 = h_2 = 5$).	39
2.9	The boundary between the external offers' accept and reject regions for different values of selling price p_e and requested service level (γ) in independent operation ($\lambda_1 = 3, \lambda_2 = 4, \mu_1 = \mu_2 = 5, \lambda_{e1} = \lambda_{e2} = 4, p_1 = p_2 = 100, h_1 = h_2 = 5$).	40
2.10	The boundary between the external offers' accept and reject regions for different values of selling price p_e and requested service level γ in static and dynamic networks with two heterogeneous members ($\mu_1 = \mu_2 = 5, \lambda_{e1} = \lambda_{e2} = 4, p_1 = p_2 = 100, h_1 = h_2 = 5$).	41
2.11	The percentage of profit increase ($\Delta\%$) for different values of selling price p_e and requested service level γ in a network with dynamic routing policy and two heterogeneous members ($\lambda_1 = 3, \lambda_2 = 4, \mu_1 = \mu_2 = 5, \lambda_{e1} = \lambda_{e2} = 4, p_1 = p_2 = 100, h_1 = h_2 = 5$).	41
2.12	External customers acceptance and rejection areas for network based operations ($\mu = 5, p = 100, h = 5, N = 10$).	44

2.13	Optimal behavior of the system for different number of members in the network ($\lambda = 2, \lambda_e = 2, \mu = 5, p = 100, p_e = 70, h = 0.05, \gamma = 0.8$). . .	45
2.14	Optimal behavior of the system for different number of members in the network ($\lambda = 2, \lambda_e = 6, \mu = 5, p = 100, p_e = 70, h = 0.05, \gamma = 0.8$). . .	46
2.15	Network based service level and independent producer based service level for different values of price ($\mu = 5, \lambda_{e_i} = 4, p = 100, h = 5$). . . .	47
2.16	(a) Percentage of profit increase for cooperative network compared to independent operation for different values of price ratio p/p_e (b) Percentage of profit increase for cooperative network compared to independent operation for different values of price p_e ($\lambda = 4, \mu = 5, \lambda_e = 20, p = 100, h = 5, \gamma = 0.85, N = 10$).	47
2.17	Profit increase $\Delta\%$ for different values of price p_e and requested service level γ ($\lambda = 4, \mu = 5, \lambda_{e_i} = 2, p = 100, h = 5, N = 10$).	48
3.1	State transition diagram of the system with two stations and quality modeling	60
3.2	Sample realization for the system with multiple up and down states ($\mu_u = 1.2, \mu_d = 1, p = 0.01, r = 0.1, p' = 0.05, r' = 0.10, g = 0.05, h = 0.10, r_Q = 0.10$)	61
3.3	Effect of first stage processing rate on production rate and expected inventory level in Markovian model with multiple up and down states ($\mu_d = 1, p = 0.005, r = 0.15, p' = 0.015, r' = 0.15, g = 0.01, h = 0.2, r_Q = 0.15, Z = 17$)	74
3.4	Effect of second stage processing rate on production rate and expected inventory level in Markovian model with multiple up and down states ($\mu_u = 1, p = 0.005, r = 0.15, p' = 0.015, r' = 0.15, g = 0.01, h = 0.2, r_Q = 0.15, Z = 17$)	75

3.5	Effect of coefficient of variation on expected inventory level in general model with Chi-square distribution for operational and failure times ($\mu_u = 1.5, \mu_d = 1, Z = 20$)	76
3.6	Effect of coefficient of variation on production rate in general model with Chi-square distribution for operational and failure times ($\mu_u = 1.5, \mu_d = 1, Z = 20$)	76
3.7	Optimal values of buffer capacity for specified production rates in the model with multiple up and down states and exponential transition times ($\mu_u = 2, \mu_d = 1, p = 0.1, r = 0.15, p' = 0.1, r' = 0.15, g = 0.1, h = 0.2, r_Q = 0.15$)	77
3.8	Optimal values of production rate and cycle time in the model with multiple up and down states and exponential transition times ($\mu_u = 2, \mu_d = 1, p = 0.1, r = 0.15, p' = 0.1, r' = 0.15, g = 0.1, h = 0.2, r_Q = 0.15, c_p = 50$)	78
3.9	Convergence of the simulated production rate to the exact value in the model with multiple up and down states and exponential transition times ($\mu_u = 1, \mu_d = 1, p = 0.01, r = 0.1, p' = 0.05, r' = 0.1, g = 0.05, h = 0.1, r_Q = 0.1, Z = 10$)	78
4.1	Telemedicine platform	89
4.2	The state space diagram of queueing model for doctor's operation . .	94
4.3	The number of participating doctors in the telemedicine platform (N) for different values of doctors' utilization (ρ) and platform desired service level (γ_d). ($\mu = 1, \lambda_t = 25$)	98
4.4	The utilization values of participating doctors in the telemedicine platform (ρ) for different numbers of members (N) and platform desired service level (γ_d). ($\mu = 1, \lambda_t = 25$)	99

4.5	The offered service level for the platform patients by each doctor in the telemedicine platform, γ_p and the effective arrival rate of platform patients, $\lambda_e(\mu = 1, \lambda_t = 25)$	100
4.6	The effective arrival rate of platform patients for each doctor in the network based on the doctors' utilization and their offered service level for the platform patients. ($\mu = 1$)	100
4.7	The number of participating number of doctors in the platform for different values of platform patients request rate and for three different values of doctors' utilization ($\mu = 1, \gamma_d = 0.9$)	101
4.8	The platform offered service level for different values of platform patients request rate and for three different numbers of participating members, N ($\mu = 1, \rho = 0.6$)	102

NOMENCLATURE

MILP: Mixed integer linear program

JSQ: Join the shortest queue

SQA: Single queue approximation

MDP: Markov decision process

FCFS: First come first serve

Chapter 1

INTRODUCTION

In response to the economic condition and competition in the business environment, manufacturers and service providers began using cooperation-based business models to utilize their capacity effectively and access more customers. Different mechanisms enable the producers or the service providers to work together and share resources. These mechanisms facilitate cooperation among different members and provide more efficient services at a lower cost. Sometimes these mechanisms are implemented through a platform between the customers and cooperating members. Cooperation may allow the members to benefit from cost advantages due to economies of scale effects, utilization improvements due to resource pooling, and flexibility enrichments among other benefits.

This dissertation aims at modeling and analysis of cooperative production and service networks. These stochastic models allow us to understand the operation of the networks and design these networks in an effective way. The dissertation is structured in three parts. In the first part, we present a detailed operational model of a cooperative manufacturing network among a number of producers and discuss the benefits of this kind of structure based on the system parameters. In the second part, we provide a mathematical programming approach to perform joint simulation and optimization of general continuous flow production systems. Finally, the last part is about the modeling and analysis of cooperative service network which enables us to examine the network designing problem.

In outline, the first and second parts are both about the stochastic models of the production systems but the second part could be considered as a separate study which

focuses on the simulation of the system's dynamics using mathematical programming framework. The first and the last part of this dissertation present a detailed investigation of the cooperation-based stochastic models in manufacturing and service systems, respectively. In the following paragraphs, we give a brief summary of the research models which are developed in these parts and the main contributions.

We dedicate the first part of this dissertation to the modeling and analysis of a cooperative production network. In this part, we examine the cooperative production business model for a group of producers serving their own customers and also have access to external customers who can make an agreement to buy products at a lower price if a desired service level can be guaranteed. Since the producers may not meet the desired service level requirement of the external customers at the offered price on their own, they participate in a cooperative network. The cooperative network consolidates the external customers for its members and routes an arriving external customer to one of the participants based on the network routing policy.

It is expected that connecting a wide range of producers in a cooperative network will allow the network to allocate the orders dynamically among the members and make it possible to fulfill orders that were not economically viable when the producers work independently. Hence, developing business models for cooperative networks is essential to realize the expected benefits of cooperation. These models enable us to describe how manufacturing services provided and how they can be deployed and paid in network structure (Putnik, 2012; Wu et al, 2013; Ren et al, 2014).

Since several manufacturing enterprises have shown great interest for creating networks in order to meet emerging market needs and business opportunities, manufacturing networks have become an attractive topic for the researchers in the recent years. Different terms such as *Dynamic Manufacturing Networks* (Papakostas et al, 2013; Ferreira et al, 2014; Ivanov et al, 2016) and *Virtual Manufacturing Networks* (Wadhwa et al, 2009) have been also used in addition to *Collaborative Manufacturing Networks* Danilovic and Winroth (2005).

Our main objective in this part of work is to answer the following main question:

What is the benefit of joining a cooperative network depending on the characteristics of the producers and the external customers? In order to answer this question, we also answer the following questions regarding the operation of this cooperative network: How do the producers manage their production and decide on meeting the demand from different groups of customers? How does the network operate depending on different levels of information exchange between the producers and the network? How do the producers decide on joining the cooperative network or not?

Depending on the level of information that the participating manufacturers share through their cooperation, we examine the operation of the network under the static and dynamic routing policies. For the network operating with the static routing policy, we determine the optimal shares of the arriving demand to be allocated to each member. For dynamic network, we determine the optimal production and rationing policies for each participating manufacturer as well as the optimal routing policy for the network by using linear programming approach to solve the resulting Markov decision process model. Stock allocation and production problems have been analyzed in limited capacity production systems by several researchers. Ha (1997b), Véricourt et al (2001), Véricourt et al (2002), and Gayon et al (2009b) are some examples of these papers in the literature. Different from these works, we combine the production and rationing decisions of the producers and the routing decision of the network in the same cooperation based model. We also present structural properties of the routing policies and a numerical study that compares the performances of both kinds of routing strategies in different kinds of network with homogeneous and heterogeneous participants.

We conclude that a cooperative network can provide financial benefits for the producers by better utilizing their capacity and serving more external customers. It also allows the external customers to obtain goods at a lower price with the desired service level. All these benefits are highly dependent on the parameters of the producers, the parameters of the external market and the level of information exchange between the network and its members.

In terms of reducing the computational effort, we developed an accurate approximation method to analyze the system with a high number of homogeneous producers which shows very high accuracy in compare to the exact solution. According to our approximate analysis, it is always beneficial for homogeneous manufacturers to join the cooperative network.

While there are recent papers on production networks, virtual enterprises and cloud manufacturing, none of these studies uses a detailed operational model to discuss cooperation-based business models. We consider showing the benefits of a cooperation-based business model for a cooperative network by analyzing a detailed operational model and examining optimal control policies jointly under the service level constraint as the main contribution of the first part.

In the second part of the dissertation, we use mathematical programming approach for simulation and optimization of a general continuous flow production system with an intermediate finite buffer. In this system, we modeled each station by a discrete state space-continuous time process with given transition time distributions between the states and a set of flow rates associated with each discrete state.

The mathematical programming approach has been proposed for joint simulation and optimization of discrete event systems with continuous time and discrete state space (Schruben, 2000). Matta (2008) used a mathematical programming representation for optimization based simulation of discrete material flow production systems. Alfieri and Matta (2012a) and Alfieri and Matta (2012b) conducted two studies investigating approximate simulation of multistage production systems and pull-controlled single-product serial discrete flow manufacturing systems. Schwarz and Stolletz (2013) analyzed time-dependent stochastic flow lines using mathematical programming. This study extends the mathematical programming approach that has been developed to analyze discrete materials flow systems for the general continuous material flow systems.

We generalize the previous approach developed by Tan (2015) for a specific system with only two states for a station and deterministic demand to a two-stage system

with multiple states with different flow rates for each station. Since the method is based on general description of stations, a wide range of continuous flow production systems can be analyzed with this method.

We develop a set of equations to determine the critical time instances of the sample trajectory of the buffer that correspond to state transitions, buffer dynamics and changing flow rates. Embedding these equations into a mixed integer linear programming (MILP) formulation enables us to perform joint simulation and optimization of the system. Our numerical results show a significant computational improvement compared to using a discrete-event simulation. As a result, the mathematical programming approach is proposed as a viable alternative method for performance evaluation and optimization of continuous-flow systems with a finite buffer.

There are three main contributions of the second part: First, we present a computationally efficient mathematical programming approach to analyze and optimize a general two-station continuous flow system with a finite buffer and generally distributed state transition times. Second, the mathematical programming representation is analyzed and compared with the direct solution of the set of equations that represents the system dynamics in a formal way. More specifically, it is shown that the solution obtained by the mathematical program is the unique solution of the set of equations that describe the system performance. Furthermore, the mathematical programming representation is used to analyze the properties of the system performance in an alternative way. The monotonicity property of the dependence of the production rate on the buffer level that is a known property of production systems is shown by using the mathematical programming representation. Finally, the computational performance of the proposed method is analyzed in detail. It is shown that the computational performance depends on the rates of change among different system states, and it does not depend on the number of states at each stage and on the buffer capacity.

We use mathematical programming approach as an efficient framework to develop analytical models and analyze the stochastic operation of two manufacturing structures in the first two parts of this dissertation. In line with the first part, in the third part

of the dissertation, we consider the operation of a cooperative service network. One example of this kind of structure is telemedicine platform. Motivated by the rapid propagation of telemedicine in the health care systems, we present a stochastic model to examine the operation and design of a telemedicine platform in this part.

It is very important to consider the effective design and management of telemedicine platform to offer the patients a desired service level, increase the utilization of the resources and maximize the profit. Much of the published works in the literature concentrates on the power of the technology and the content and quality of interactions it supports in telemedicine services (Hung and Zhang, 2003; Niyato et al, 2007; Devaraj and Ezra, 2011; Warren et al, 2011). However, there are a few number of works which focus on the operations management in telemedicine context (Rajan et al, 2018; Erdogan et al, 2018) and designing an efficient platform (Körpeoğlu et al, 2014). To fill this gap, we consider a detailed operational analysis of a telemedicine platform when it is formed as a cooperative service network and discuss the designing issues of this structure to obtain the benefits of the cooperation both for the patients and the doctors.

The model contains a number of physicians who receive their own patients in person. Collaboration of these doctors in telemedicine platform is expected to provide remote assistance for the patients from one site to another, and therefore, generates an additional stream of patients for the doctors to achieve their full potential in serving the patients. The operation of the whole platform is analyzed approximately by using single queue approximation method. Based on our results, we consider platform design problem where the decision variable is the number of doctors in the platform, and discuss the problem for different utilization levels, the patients' arrival rates and the desired service levels.

Hence, our main objective is to answer the following main question: How do the designers of telemedicine platform decide on different parameters of the telemedicine cooperative network? and to answer this question, we should also answer the following question: How do the doctors decide whether to become available in order to receive

the patients from telemedicine platform when their own patients are already waiting in the system? This question is important due to non-preemptive characteristic of the model.

In different kinds of applications, a single server is usually shared by multiple types of customers with different revenue, arrival process and service time. So, there is a vast literature on defining the optimal dynamic admission control for revenue management in service and manufacturing systems. Lippman and Ross (1971), Harrison (1975), Örmeci et al (2001), Örmeci and van der Wal (2006), Ulukus et al (2011), and Turhan et al (2012) are some examples of these studies.

To solve the dynamic admission problem, we model the operation of doctors using Markov decision process and derive some properties of the value function, based on which we further prove the existence of threshold type optimal control policy. The threshold type policy prescribes the optimal solution to achieve the maximum discounted profit criterion over a finite time horizon. We also perform the stationary analysis of the system and show that the optimal policy is a strict non-preemptive priority for the doctors' own patients to maximize total average profit over an infinite time horizon. Our numerical experiments contain a full discussion about the designing parameters of the platform to assure a desired user experience while leading to the platform and its member's profitability.

The main contribution of this part is two-fold. Firstly, we present a detailed operational model of the cooperative service network in the telemedicine platform setting and discuss the designing issues to obtain the efficient performance. Secondly, we prove that the optimal policy of a server that faces the demand from two different customer streams, one that join the queue and the other one that is lost, is a non-preemptive threshold type policy. None of these issues have been addressed in the literature before.

The rest of this dissertation is organized as follows: In Chapter 2, we consider modeling and analysis of a cooperative production network. Chapter 3 is about simulation and optimization of continuous flow production systems by using mathe-

mational programming. Chapter 4 addresses modeling and analysis of a cooperative service network with application in telemedicine platform. An overall conclusion of the dissertation and future works are provided in Chapter 5. Additionally, some necessary proofs are given in Chapter 6.



Chapter 2

MODELING AND ANALYSIS OF A COOPERATIVE PRODUCTION NETWORK

Since several manufacturing enterprises have shown great interest for creating networks in order to meet emerging market needs and business opportunities, manufacturing networks have become an attractive topic for the researchers in the recent years. Different terms such as *Dynamic Manufacturing Networks* (Papakostas et al, 2013; Ferreira et al, 2014; Ivanov et al, 2016) and *Virtual Manufacturing Networks* (Wadhwa et al, 2009) have been also used in addition to *Collaborative Manufacturing Networks* (Danilovic and Winroth, 2005).

With the advances in technology, the producers can establish these networks easily. The *Cloud Manufacturing* platforms allow producers to cooperate and interact with their customers in a more effective way. The *Cloud Manufacturing* is defined as a manufacturing model that exploits on-demand access to a shared collection of distributed manufacturing resources to form temporary production lines which enhance efficiency and allow for optimal resource loading in response to variable customer's demand (Li et al, 2010; Wu et al, 2013; Zhang et al, 2014). It is expected that connecting a wide range of producers in a cloud manufacturing network will allow the network to allocate the orders dynamically among the members and make it possible to fulfill orders that were not economically viable when the producers work independently. Developing business models that describe how manufacturing services provided and how they can be deployed and paid is important to realize the expected benefits of cooperation (Putnik, 2012; Wu et al, 2013; Ren et al, 2014). An example

for this kind of setting is MFG.com which is one of the largest contract manufacturing marketplaces and enables the customers to access manufacturers through an easy to use on-line platform.

Our study is motivated from such a situation that a group of producers can cooperate in a network setting. These manufacturers produce a single product for their own inventories and serve their own customers. They also have access to another group of customers, external customers, who offer a lower price for the products compared to the producers' own customers and also require a specific service level. For example, a producer may want to use his excess capacity to produce for other companies as a private label supplier. In order to be accepted as a private label supplier, he should meet the service level requirements of the company that he will produce for. If a producers can meet the service level requirement of external customers in an economically viable way, these customers bring additional business and allow the producer to utilize his resource more effectively. However, due to the service level requirements and also the lower price, these customers are not usually served by individual producers.

The cooperative network we analyze consolidates the orders from the external customers and routes the orders to available producers in such a way that the desired service level requirement of the external customers is met. Collaboration of producers in this network is expected to meet the desired service level at a lower cost and therefore generates additional business to its members and utilizes their resources more effectively.

In this paper, our main objective is to answer the following main question: *What is the benefit of joining a cooperative network depending on the characteristics of the producers and the external customers?* In order to answer this question, we also answer the following questions regarding the operation of this cooperative network: *How do the producers manage their production and decide on meeting the demand from different groups of customers?; How does the network operate depending on different levels of information exchange between the producers and the network?; and How do*

the producers decide on joining the cooperative network or not?

We model the system with exponentially distributed production times and exponential order inter-arrival times as a Continuous-time Markov chain to evaluate the performance of the system when the production and sales decisions of the producers and the routing decision of the network are given optimally. We analyze the operation of the network in two different information settings. When the producers do not share any information about their real-time finished goods' inventory levels with the network, the network uses a static routing policy to allocate the external traffic among its members in an optimal way. When the information about inventory levels are shared with the network continuously, the network uses a dynamic optimal routing policy to send an arriving order to one of the producers in the network. We analyze the independent operation and the network operation with static routing policy analytically. For the network operation with the dynamic routing policy and heterogeneous producers, we present a linear programming formulation to determine the optimal production, sales, and routing policies. Moreover, for the network operation with dynamic routing policy and homogeneous producers, we develop an approximation method that allows us to analyze the cooperation network with a high number of members. We determine whether it is beneficial for a producer to join the cooperative network by comparing the expected profit of operating independently with the expected profit of participating in the cooperative network. We show that the cooperative network can increase the profits of the producers significantly for the high values of requested service level.

2.1 Past Work

Our study is related to the production control with multiple demand classes literature and also to the cooperation in supply chain literature.

2.1.1 Production Control with Multiple Demand Classes

Whenever the limited capacity of a supplier is to be shared among multiple groups of customers with different profit margins, it may be more profitable for the supplier to

reject a group of customers in order to reserve the available inventory for upcoming more profitable customers. In addition, if the production capacity is limited, the producer should decide on how to control production and how to serve different demand classes.

One of the earliest works in this area is presented by Topkis (1968), who developed an optimal stock rationing problem for an uncapacitated system. This research has been extended with different kinds of assumption in various directions, e.g. Nahmias and Demmy (1981), Cohen et al (1988), and Frank et al (2003). Deshpande et al (2003) give a detailed review of this literature.

Queuing-based models have been used to analyze stock allocation problems in limited capacity production systems. Ha (1997a) analyzed the optimal rationing and production policy of a lost-sale manufacturing system by modeling the production system as a multi-class $M/M/1$ make-to-stock queue and showed that it is optimal to satisfy an arriving demand when the level of inventory is above the threshold level for that class of customers and to reject them otherwise. Ha (1997b) also analyzed a system with two different classes of customers and backorder. Véricourt et al (2002) generalized this problem for multiple classes of customers and specified that it is optimal to satisfy an arriving demand from a customer if the on-hand inventory is above the threshold for that customer and to backorder the demand otherwise. Véricourt et al (2001) also performed a comparison among three general kinds of control policies for a make-to-stock production system with multiple classes of customers and showed that the multi-level rationing policy always outperforms the other two policies. During the recent years, the rationing policies have been investigated for various systems with different processing time distributions (Lee and Hong, 2003; Gayon et al, 2009b), with batch ordering (Huang and Iravani, 2007; Pang et al, 2014), with batch arrivals (Çil et al, 2009), among others (Carr and Duenyas, 2000; Duran et al, 2007; Gayon et al, 2009a; Liu et al, 2015).

The main contributions of this study to this literature is combining production and rationing decisions of the producer and the routing decision of the network in the

same model. We consider analyzing the optimal production, rationing, and routing decisions jointly and showing the structure of the optimal policies computationally.

2.1.2 Cooperation on Production, Inventory, and Sales

The benefits of pooling capacity and inventory had been shown in different studies. Benjaafar et al (2005) examined inventory pooling in production-inventory systems and showed that utilization, demand and process variability, control policy, service levels, and the structure of the production process all play a role in determining precisely how valuable pooling might be. Yu et al (2015) presented models to study the benefit of capacity pooling among independent service firms by modeling facilities as $M/M/1$ queues and determining capacity by the service rate. The authors showed that capacity pooling reduces total cost and all firms can benefit under an appropriate cost sharing scheme which is always feasible. Tan (2006) presents a stochastic model to analyze cooperation of producers on production capacity. An auction based allocation method is proposed to distribute large orders and additional benefits among independent producers. Demand allocation problem in a system with multiple inventory locations is also studied by Benjaafar et al (2004) and Benjaafar et al (2008). Inventory sharing through transshipment has also attracted a great deal of attention from researchers in the recent years (Zhao et al, 2006; Yang and Qin, 2007; Zhao et al, 2008). In this line of research, the emphasis has been on the capacity and inventory management models proposed for the pooled inventory.

In addition to cooperation on capacity and pooling inventory, companies can also cooperate in marketing and sales activities. Through marketing cooperation, firms reduce marketing costs by bundling their products. Using empirical data, Bucklin and Sengupta (1993) identify situations in which marketing cooperation are most successful and Swaminathan and Moorman (2009) explore whether such cooperation indeed add value to the firms. Assortment-based cooperation is a cooperation scheme that spans inventory pooling. Firms improve the utilization of their production capacities through additional demand streams and need to satisfy the demand of the

other firm, essentially treating their own stock as pooled inventory. Akçay and Tan (2008) examine an assortment-based cooperation among independent firms and show that such a cooperation scheme between symmetric single-product firms is always beneficial, whereas a threshold-type criterion should be satisfied so that assortment-based cooperation is beneficial for asymmetric firms. However, their model inherently assumes that firms do not have any capacity limitations. Tan and Akçay (2014) extends the model in Akçay and Tan (2008) to capture the effects of capacity limitations on benefits of assortment-based cooperation in a stochastic setting.

We consider showing the benefits of a cooperation-based business model for a cooperative network by analyzing a detailed operational model and examining optimal control policies jointly as the main contribution of this study. We also developed an accurate approximation method to analyze the system with a high number of homogeneous producers which reduces the computational effort very much and shows very high accuracy in compare to the exact solution.

2.2 Model Description

We consider N production facilities that produce a single product to their stocks and face with two different groups of customers.

Production Facilities. The production time for Producer i is exponentially distributed with mean μ_i^{-1} , $i = \{1, \dots, N\}$. When a demand arrives, it is either satisfied immediately from the on-hand inventory or is rejected and lost. The service level is defined based on the fill rate. The corresponding unit inventory holding cost for each producer is h_i . The production costs are set to zero by scaling the prices accordingly.

Customer Demand. The first group of customers is referred as the manufacturers' own customers. Producer i 's own customers arrive according to a Poisson process with an arrival rate of λ_i and pay a price of p_i . The requests of the manufacturers' own customers are satisfied by them directly.

The second group is referred as the external customers. We assume that the demand of the external customers for all producers are generated from the same class

of customers. In the private label production example, these are the companies that would like to source private-label products from various suppliers. They expect a desired service level (fill rate) of γ from the manufacturers.

Producer i 's external customers arrive according to a Poisson process with rate λ_{ei} . They pay a lower price of p_{ei} , $p_i > p_{ei}$, and the business offer introduced by the external customers is $(\lambda_{ei}, p_{ei}, \gamma)$. The manufacturers accept the offer from this group of customers only if they can provide the desired service level in a profitable way. Otherwise they reject the second group of customers since they cannot meet their service level expectation, and continue their operation by serving only their own customers.

Figure 2.1 depicts the independent operation of a producer. For independent setting, the producer should decide about optimal production and selling policies to maximize his profit.

Network. A producer can participate in a cooperative network with other manufacturers. The network consolidates the external customers of all the members and serve them by routing an arriving external demand originated from the external customers of one of the members to one of the producers in the network. The total arrival rate of the external customers to the network is $\lambda_e = \sum_{i=1}^N \lambda_{ei}$. Please note that in the network operation, each member continues to receive his own customers demand independently of the network.

Routing Policy of the Network Based on Information Setting. Based on the level of information that is shared by the producers, we analyze the operation under two different kinds of routing policies: the static routing policy where the real-time information about inventory levels is not shared with the network and the dynamic routing policy where the producers share their inventory levels continuously with the network. For both cases, the producers share their capacity, price, cost, and demand information with the network. In the static routing case, the operation is optimized to define the share of each producer from the external customers, base stock and rationing levels. In the dynamic case the optimal production, rationing and

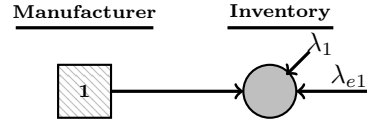


Figure 2.1: Accepting external customers directly

routing policies should be determined to understand the operation of the network.

The parameters for the independent (no cooperation) and cooperation under the static and the dynamic networks are differentiated by using subscript j , $j \in \{nc, cs, cd\}$, where nc stands for the independent (no cooperation case), cs stands for cooperation in the cooperative network under the static routing policy and cd stands for cooperation in the cooperative network under the dynamic routing policy. Accordingly, the inventory level for Producer i operating independently is defined by I_{nc_i} . In the network based operation, the inventory level of Producer i is I_{cs_i} and I_{cd_i} under the static and the dynamic routing policies respectively. Figure 4.1 shows the schematic of the cooperative network. For the network setting, each participating producer should decide about the optimal production and selling policies. The network should also decide about optimal routing policy of external customers among members.

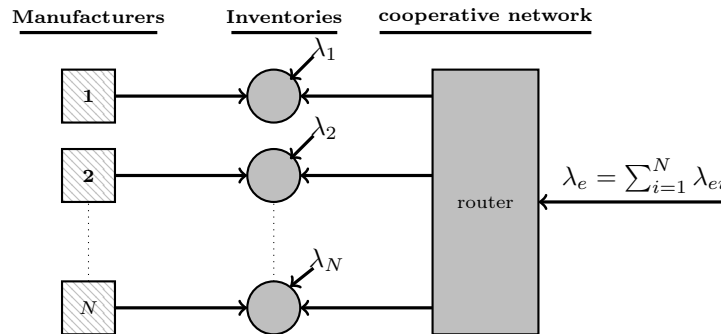


Figure 2.2: Cooperative network for external sales

2.3 Independent Operation (No cooperation case)

When a production system with limited capacity serves two different classes of customers with different arrival rates and profit margins, it may be more profitable for the producer to reject a less profitable customers to keep the available inventory for more profitable customers.

Since there is a service level constraint for the external demand in our setting, an independent producer should make a decision to accept or reject the business offer $(\lambda_{ei}, p_{ei}, \gamma)$ from this group of customers. When the business offer is accepted, the producer should provide the desired service level for the external customers. This may lead to a profit decrease by losing some of the demand from the producer's own customers who bring more profit. In this case, the producer rejects the external customers business offer and sells the products only to his own customers. However, if the additional profit obtained from the external customers compensates the loss due to losing some of the demand from the producer's own customers and brings more profit, this business offer is accepted by the producer. We determine the optimal decision to accept or reject the business offer from external customers by comparing the expected profits for both cases.

Ha (1997a) studied a single-item production system with several demand classes and lost sales. For the discounted cost minimization objective of the problem over an infinite time horizon, it is shown that the optimal production policy is base stock policy. In addition, for each class of demand, there is an admission threshold level at or below which it is optimal to start rejecting this class of demand in anticipation of future arrival of more profitable demands.

Accordingly, Producer i operates with a base-stock level (S_{nc_i}) to decide whether to produce, and with a rationing level (R_{nc_i}) to decide whether to satisfy an arriving external demand.

Whenever Producer i accepts external customers, the profit rate function is defined

as follows:

$$\Pi_{nc_i}^{accept} = \lambda_i(Pr(I_{nc_i} > 0))p_i + \lambda_{ei}(Pr(I_{nc_i} > R_{nc_i}))p_{ei} - h_i\bar{I}_{nc_i}. \quad (2.1)$$

In this equation, the first term shows that Producer i makes a revenue of p_i from an arrival of its own customers as long as the finished goods inventory is not empty, i.e., there is at least one part in the inventory. $Pr(I_{nc_i} > 0)$ is the stationary probability that the inventory level I_{nc_i} is greater than zero. The second term shows that whenever the inventory level is greater than the determined rationing level, it accepts an additional demand from the external customers and makes a revenue of p_{ei} per unit. $Pr(I_{nc_i} > R_{nc_i})$ is the stationary probability that the inventory level I_{nc_i} is greater than the rationing level R_{nc_i} . Finally, the last term depicts the inventory holding cost. $\bar{I}_{nc_i}$ is the expected level of inventory for producer i .

The traffic intensity at producer i due to his own customers is defined by $\rho_i = \lambda_i/\mu_i$. Also, whenever Producer i accepts the external business offer, the traffic intensity due to the all customers (own customers and external customers) is given by $\xi_{nc_i} = (\lambda_i + \lambda_{ei})/\mu_i$. The performance measures of the system are determined from the steady-state probabilities as follows:

$$Pr(I_{nc_i} > R_{nc_i}) = 1 - \frac{(1 - \xi_{nc_i})\xi_{nc_i}^{S_{nc_i}-R_{nc_i}}(1 - \rho_i^{R_{nc_i}+1})}{(1 - \rho_i)(1 - \xi_{nc_i}^{S_{nc_i}-R_{nc_i}}) + (1 - \xi_{nc_i})\xi_{nc_i}^{S_{nc_i}-R_{nc_i}}(1 - \rho_i^{R_{nc_i}+1})}, \quad (2.2)$$

$$Pr(I_{nc_i} > 0) = 1 - \frac{(1 - \xi_{nc_i})(1 - \rho_i)\xi_{nc_i}^{S_{nc_i}-R_{nc_i}}\rho_i^{R_{nc_i}}}{(1 - \rho_i)(1 - \xi_{nc_i}^{S_{nc_i}-R_{nc_i}}) + (1 - \xi_{nc_i})\xi_{nc_i}^{S_{nc_i}-R_{nc_i}}(1 - \rho_i^{R_{nc_i}+1})}, \quad (2.3)$$

$$\begin{aligned} \bar{I}_{nc_i} = E[I_{nc_i}(t)] = & S_{nc_i} - \frac{(\xi_{nc_i}(1 - \rho_i)/(1 - \xi_{nc_i}))(1 - \xi_{nc_i}^{S_{nc_i}-R_{nc_i}} - (1 - \xi_{nc_i})(S_{nc_i} - R_{nc_i})\xi_{nc_i}^{S_{nc_i}-R_{nc_i}-1})}{(1 - \rho_i)(1 - \xi_{nc_i}^{S_{nc_i}-R_{nc_i}}) + (1 - \xi_{nc_i})\xi_{nc_i}^{S_{nc_i}-R_{nc_i}}(1 - \rho_i^{R_{nc_i}+1})} \\ & - \frac{\rho_i(1 - \xi_{nc_i})/(1 - \rho_i)}{(1 - \rho_i)(1 - \xi_{nc_i}^{S_{nc_i}-R_{nc_i}}) + (1 - \xi_{nc_i})\xi_{nc_i}^{S_{nc_i}-R_{nc_i}}(1 - \rho_i^{R_{nc_i}+1})} \\ & \times \frac{(\xi_{nc_i}/\rho_i)^{S_{nc_i}-R_{nc_i}}(\rho_i^{S_{nc_i}-R_{nc_i}} - \rho_i^{S_{nc_i}+1} + (1 - \rho_i)[(S_{nc_i} - R_{nc_i})\rho_i^{S_{nc_i}-R_{nc_i}-1} - (S_{nc_i} + 1)\rho_i^{S_{nc_i}}])}{(1 - \rho_i)(1 - \xi_{nc_i}^{S_{nc_i}-R_{nc_i}}) + (1 - \xi_{nc_i})\xi_{nc_i}^{S_{nc_i}-R_{nc_i}}(1 - \rho_i^{R_{nc_i}+1})}. \end{aligned} \quad (2.4)$$

Please note that Equations (2.2), (2.3) and (2.4) are true when $\rho_i \neq 1$ and $\xi_{nc_i} \neq 1$.

The equations for $\rho_i = 1$ or $\xi_{nc_i} = 1$ are given in Appendix.

The optimal profit rate of Producer i when it accepts the business offer from the external customers, $\Pi_{nc_i}^{accept*}$, is obtained by maximizing Equation (2.1) to determine the optimal base stock level $S_{nc_i}^*$ and the rationing value $R_{nc_i}^*$. The maximization problem is subject to the external customers' requested service level constraint as follows:

$$\begin{aligned} \Pi_{nc_i}^{accept*} &= \max_{S_{nc_i}, R_{nc_i}} \Pi_{nc_i}^{accept} \\ \text{subject to:} \quad &\gamma_{nc_i} \geq \gamma. \end{aligned} \quad (2.5)$$

where $\gamma_{nc_i} = Pr(I_{nc_i} > R_{nc_i})$ is the offered service level for external customers.

Whenever the producer rejects the external business offer, his profit is the difference between the revenue obtained from his own customers and the inventory carrying cost. The expected profit rate function is given as follows:

$$\Pi_{nc_i}^{reject} = \lambda_i (Pr(I_{nc_i} > 0)) p_i - h_i \bar{I}_{nc_i}. \quad (2.6)$$

The producer chooses the base-stock level S_{nc_i} to maximize his profit. Therefore,

$$\Pi_{nc_i}^{reject*} = \max_{S_{nc_i}} \Pi_{nc_i}^{reject}. \quad (2.7)$$

When the external business offer is rejected, we set $S_{nc_i} = R_{nc_i}$ and $\rho_i = \xi_{nc_i} = \lambda_i / \mu_i$ in Equations (2.2) - (2.4). The resulting formulations in this case are given in the Appendix.

Based on Equations (2.5) and (2.7), the optimal profit rate function for the independent operation is given as

$$\Pi_{nc_i}^* = \max\{\Pi_{nc_i}^{accept*}, \Pi_{nc_i}^{reject*}\}. \quad (2.8)$$

The optimal decision of Producer i to accept or reject the external business offer when it operates independently, $g_{nc_i}^*$, is determined by comparing the expected profit rates of both cases:

$$g_{nc_i}^* = \begin{cases} \text{accept} & \text{if } \Pi_{nc_i}^{accept*} \geq \Pi_{nc_i}^{reject*} \\ \text{reject} & \text{if } \Pi_{nc_i}^{accept*} < \Pi_{nc_i}^{reject*} \end{cases}. \quad (2.9)$$

There might be three reasons for an independent producer to reject an external business offer: The first one is the high utilization of the producer due to his own customers (ρ). The second reason is the much higher arrival rate of the external customers compared to the production rate. Finally, if the sale price of the external customers is much lower than the price of the producer's own customers, the producer decides reserving more inventory for its own customers due to their higher profit margin. Since there is a service level requirement, the producer will reject the external business offer under a certain selling price defined according to the system parameters.

Figure 2.3 shows the boundary between the accept and reject regions for two independent producers evaluating the external business offers (λ_{ei} , p_{ei} , γ). Each producer has a different utilization level and considers the external business offers for three different values of λ_e . For each value of λ_e and selling price p_e , the independent producers can offer external customers a maximum service level and accept them only if their requested service level is less than or equal to this maximum value. Therefore, all the points over each boundary line with the specified value of λ_e in Figure 2.3, defines the region where the producer cannot provide the desired service level for the external customers and prefers to reject them. Consequently, all the points under each boundary line defines the points where the producer is able to provide the desired service level and accepts external business offer. Comparing both graphs in Figure 2.3(a) and Figure 2.3(b) shows that the accept region is much larger for the producer depicted on the left with a lower level of utilization ($\rho = 0.4$). This means that the producer with a high utilization could not provide the high service level required for the external offer. The figure also shows that as the arrival rate of the external customers (λ_e) increases, the producer decides to offer them a lower service level in order to keep capacity for his own more profitable customers.

As a result, the independent producers may not be able to accept external business offers based on the system parameters. Consequently, losing the potential external offers leads to lower profitability and utilization. The cooperative network analyzed in the following sections is expected to provide a solution to this issue by providing the

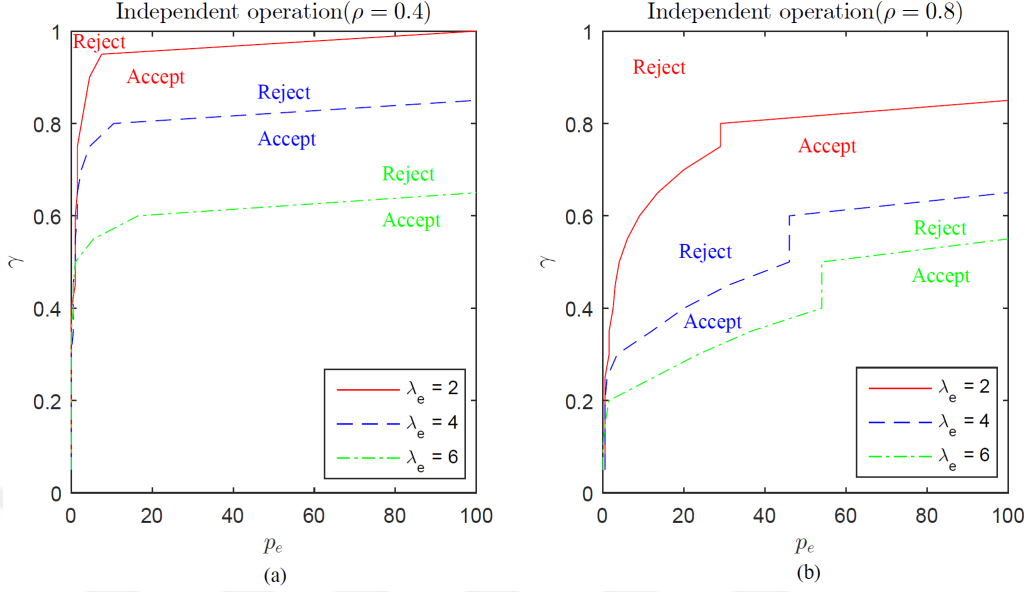


Figure 2.3: (a) The boundary between the external customers' accept and reject regions for different values of selling price p_e and requested service level γ when $\rho = 0.4$, (b) The boundary between the external customers' accept and reject regions for different values of selling price p_e and requested service level γ when $\rho = 0.8$, ($\mu = 5, p = 100, h = 5$).

ability to accept the external business offers that are rejected by the producers when they operate independently. In the following sections, we will analyze a cooperative network among N manufacturers. Based on the level of information that is shared by the producers, we analyze the operation under the static and the dynamic routing policies.

2.4 Cooperative Network with N Producers and Static Routing Policy

We first analyze the case where the network does not have on-line information about the inventory levels of the members and use a static routing policy to route the arriving external customer to one of the participants. The routing policy is optimized to define the share of each producer from the external customers, $\alpha_1, \alpha_2, \dots, \alpha_N$. Hence, α_1 is the portion of the external customer requests sent to producer 1, α_2 is the portion of these requests sent to the second producer in the network and it goes up to α_N which

is the portion of external requests sent to the producer N .

The producers decide on whether to enter the cooperation by comparing the total profit when they operate independently and when they join the cooperative network. If the cooperation brings additional profit, this surplus can be shared between the cooperating producers in such a way that each producer is better off compared to the independent operation.

Let Π_{cs} be the expected profit for the producers cooperating in a static network when member i uses the rationing threshold R_{cs_i} and the base-stock level S_{cs_i} . Then, the expected total profit rate of static network is defined as follows:

$$\Pi_{cs} = \sum_{i=1}^N \left(\lambda_i Pr(I_{cs_i} > 0) p_i + \alpha_i \lambda_e Pr(I_{cs_i} > R_{cs_i}) p_{ei} - h_i \bar{I}_{cs_i} \right) \quad (2.10)$$

The performance measures $Pr(I_{cs_i} > 0)$, $Pr(I_{cs_i} > R_{cs_i})$ and $\bar{I}_{cs_i}$ are calculated by using Equations (2.2), (2.3) and (2.4) after setting the traffic intensity due to both groups of customers as $\xi_{cs_i} = (\lambda_i + \alpha_i \lambda_e) / \mu_i$. The profit rate is maximized with respect to each producer's share of the demand from the external customers (α_i) and also the base stock (S_{cs_i}) and rationing levels (R_{cs_i}):

$$\max_{S_{cs_i}, R_{cs_i}, \alpha_i} \Pi_{cs} \quad (2.11)$$

Subject to

$$\sum_{i=1}^N \alpha_i = 1, \quad (2.12)$$

$$\sum_{i=1}^N \alpha_i \gamma_{cs_i} \geq \gamma, \quad (2.13)$$

$$0 \leq \alpha_i \leq 1 \quad \forall i \in \{1, \dots, N\}. \quad (2.14)$$

Equation (2.12) shows that all external requests are routed to the members of the network. γ_{cs_i} is the offered service level to external customers by manufacturer i participating in the network ($\gamma_{cs_i} = Pr(I_{cs_i} > R_{cs_i})$). Finally, Equation (2.13) ensures that the service level offered by the network meets the desired service level constraint of external business offer. The solution of the above optimization problem yields the

optimal shares of external demand to be allocated to each producer and gives the optimal profit. Let us assume that $g_{s_i} \in \{cs, nc\}$ defines the action of joining the cooperative network with static rationing policy (cs) or operating independently (nc). Therefore:

$$g_{s_i}^* = \begin{cases} \text{cooperation (cs)} & \text{if } \Pi_{cs}^* = \max\{\Pi_{cs}^*, \sum_{i=1}^N \Pi_{nc_i}^*\} \\ \text{no cooperation (nc)} & \text{if } \sum_{i=1}^N \Pi_{nc_i}^* = \max\{\Pi_{cs}^*, \sum_{i=1}^N \Pi_{nc_i}^*\} \end{cases}. \quad (2.15)$$

Comparing Equations (2.1) and (2.10), we can conclude that the system performance for a cooperative network with identical producers and static routing policy is the same as the independent operation of these producers. The reason for this conclusion is that the resulting optimal values of α_i will be equal to $1/N$ for all participants in the static network due to homogeneity of the members. Hence, the static network with identical members does not make any improvement for the participants. On the other hand, in a static cooperative network with heterogeneous participants, the benefits of cooperation is highly dependent on system parameters because the routing decision is optimized just according to the steady state performance and it is not sensitive to on-line inventory levels information. We will see all these outcomes in the numerical results.

2.5 Cooperative Network with N Producers and Dynamic Routing Policy

In order to analyze the operation of the cooperative network with N participating manufacturers, we determine the optimal production, rationing, and routing control policies that maximize the average profit of the system per unit time over an infinite horizon. We employ a Markov decision process framework and due to the memoryless property, the optimal control policy depends only on the current state of the system regardless of time. Thus, we denote the current state of the system by $x = [I_1, I_2, \dots, I_N]$ where $I_i \in \mathbb{I}_i = \{0, 1, \dots, M_i\}$ is the inventory level for producer $i \in \{1, 2, \dots, N\}$ participating in the network and M_i is the maximum number of products in Inventory I_i . At any time, the producers should decide on the production policy to produce or

to stop production. Let us define these two actions with 1 and 0, respectively, and the action for all members by $u = [u_1, u_2, \dots, u_N]$ where u_i is the production decision by producer i in the network. When a producer's own customer arrives, it may either be satisfied or rejected. We denote these two actions with 1 and 0 respectively and the action for all members by $f = [f_1, f_2, \dots, f_N]$ where f_i is the rationing decision of producer i for his own customers. Finally, the network should decide on the routing policy. Let us define the action by $d = \{0, 1, 2, \dots, N\}$ where 0 means rejecting the incoming external customer and i means accepting the incoming external demand and sending it to the producer i , $i = 1, 2, \dots, N$. Therefore, the control policy is described by $v = \{(u(x), f(x), d(x))\}$ and $v \in \mathbb{V}$ where $\mathbb{V}$ is the set of all admissible actions.

Given a policy $v \in \mathbb{V}$, let $n_i^o(t)$ be the number of producer i 's own customers who have been accepted by producer i up to time t and $n_i^e(t)$ be the number of external requests which is satisfied by producer i via the network channel up to time t . We seek to find the optimal control policy v that maximizes the following average profit of the system per unit time over an infinite horizon:

$$\lim_{T \rightarrow \infty} \sup_{v \in \mathbb{V}(x^0)} \left\{ \frac{1}{T} E_{x^0}^v \left[\sum_{i=1}^N \int_0^T p_i n_i^o(t) dt + \sum_{i=1}^N \int_0^T p_{ei} n_i^e(t) dt - \sum_{i=1}^N \int_0^T h_i I_i(t) dt \right] \right\}, \quad (2.16)$$

where $x^0 = [I_1^0, I_2^0, \dots, I_N^0]$ is the initial inventory level and $E_{x^0}^v$ denotes expectation over demand given an initial inventory level x^0 .

2.5.1 Exact Analysis of the Optimal Production, Rationing and Routing Policies of the Cooperative Network with N Participants Using Linear Programming Formulation

In our model, given the current state and decision, any probabilistic statement about the future of the process is completely independent of any information about the history of the process. Therefore, a linear programming formulation can be used to determine the optimal control policy for the Markov Decision Process (Bertsekas et al, 1995; Pietrabissa, 2008; Puterman, 2014; Bhandari et al, 2008; Nadar et al, 2016).

In order to develop a mathematical programming formulation, we construct a

discrete-time equivalent of the original system by using the standard tools of uniformization. To this end, we assume that the time between two consecutive transitions is exponentially distributed with rate $\psi = \lambda_e + \sum_{i=1}^N \lambda_i + \sum_{i=1}^N \mu_i$ which is the maximum transition rate of the system and rescale the time scale of the problem without loss of generality. The decision variable for the linear programming model is defined as P_{xufd} , which is the steady-state unconditional probability that the system is in state x , when decisions u, f and d are made.

Let Π_{cd} be the expected profit for the producers cooperating in a network with a dynamic routing policy. our objective is maximizing the expected long-run profit function, which is written as follows:

$$\Pi_{cd} = \sum_x \sum_u \sum_f \sum_d \Pi_{xfd} P_{xufd} \quad (2.17)$$

where

$$\Pi_{xfd} = \sum_{i=1}^N f_i \lambda_i p_i + \sum_{i=1}^{N-1} y(d, i) \lambda_e p_{ei} + w(d) \lambda_e p_{eN} - \sum_{i=1}^N I_i h_i, \quad (2.18)$$

$$y(d, i) = \frac{\prod_{\substack{j=0 \\ j \neq i}}^N (d - j)}{\prod_{\substack{j=0 \\ j \neq i}}^N (i - j)}, \quad (2.19)$$

$$w(d) = \frac{\prod_{j=0}^{N-1} (d - j)}{N!}. \quad (2.20)$$

Equation (2.18) is the profit obtained when the system is at state x and the decisions u, f and d are made. The network obtains a profit by selling the product to its own customers when $f_i = 1$ and also by selling the product to the external customer from producer $i \in \{1, 2, \dots, N\}$ when $d = i$. The last term is the inventory holding cost. Equations (2.19) and (2.20) ensure that the network gains profit by selling the product to the external customer from Producer i when $d = i$. Hence the mathematical formulation for a general dynamic network with N participating manufacturers is

given as follows:

$$\max \Pi_{cd}$$

Subject to

$$\sum_u \sum_f \sum_d \left[P_{xufd} - \frac{1}{\psi} \left\{ \sum_{i=1}^N u_i \mu_i P_{(x-e_i)ufd} + \sum_{i=1}^N f_i \lambda_i P_{(x+e_i)ufd} \right. \right. \quad (2.21)$$

$$\begin{aligned} &+ \sum_{i=1}^{N-1} y(d, i) \lambda_e P_{(x+e_i)ufd} + w(d) \lambda_e P_{(x+e_N)ufd} \\ &+ P_{xufd} \left(\psi - \sum_{i=1}^N u_i \mu_i - \sum_{i=1}^N f_i \lambda_i - \sum_{i=1}^{N-1} y(d, i) \lambda_e - w(d) \lambda_e \right) \Big\} = 0 \quad \forall x, \\ &\sum_x \sum_u \sum_f \sum_d P_{xufd} = 1 \end{aligned} \quad (2.22)$$

$$P_{xufd} = 0 \quad \forall i, I_i \in \{M_i\}, u_i \in \{1\}, \forall f, \forall d, \quad (2.23)$$

$$P_{xufd} = 0 \quad \forall i, I_i \in \{0\}, f_i \in \{1\}, \forall u, \forall d, \quad (2.24)$$

$$P_{xufd} = 0 \quad \forall i, I_i \in \{0\}, d \in \{i\}, \forall u, \forall f, \quad (2.25)$$

$$P_{xufd} = 0 \quad \forall i, I_i \in \{-1, M_i + 1\}, \forall u, \forall f, \forall d, \quad (2.26)$$

$$\sum_x \sum_u \sum_f \sum_{d=1}^N P_{xufd} \geq \gamma. \quad (2.27)$$

The first two constraints are the balance and the normalization equations, which yield the steady state probability values. Equation (2.23) ensures that Producer i stops production whenever its inventory level is equal to maximum possible value M_i . Equation (2.24) together with Equation (2.25) show that whenever the inventory is not available, it is impossible to sell an item to its own and the external customers respectively. Equation (2.26) indicates the boundary conditions and the last constraint in Equation (2.27) is to define the requested service level for external business offer. The solution of the above mathematical program yields the optimal production, rationing, and routing policies and gives the optimal profit for dynamic network, Π_{cd}^* . The optimal decision for entering the cooperative network with dynamics routing policy

(cd) or operating independently (nc), ($g_{d_i} \in \{cd, nc\}$), is also defined as follows:

$$g_{d_i}^* = \begin{cases} cd \text{ (cooperation)} & \text{if } \Pi_{cd}^* = \max\{\Pi_{cd}^*, \sum_{i=1}^N \Pi_{nc_i}^*\} \\ nc \text{ (no cooperation)} & \text{if } \sum_{i=1}^N \Pi_{nc_i}^* = \max\{\Pi_{cd}^*, \sum_{i=1}^N \Pi_{nc_i}^*\} \end{cases} \quad (2.28)$$

As an example, we will study the structural properties of the optimal production, rationing, and routing policies for a cooperative network with two producers in the following section.

2.5.2 Structure of the Optimal Production, Rationing and Routing Control Policies for the Network with Two Participants

Figure 2.4 represents the structure of optimal control policies of two different dynamic networks obtained by solving the linear program presented in Section 2.5. The left

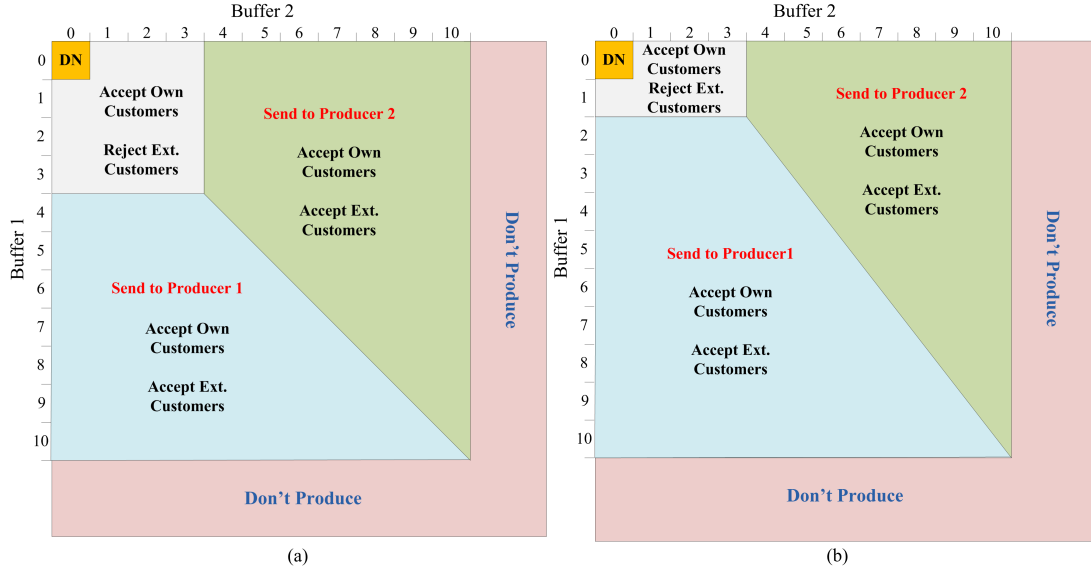


Figure 2.4: (a) Optimal production, rationing, and routing policy for a network with two homogeneous manufacturers participating ($\lambda_1 = 2, \lambda_2 = 2, \mu_1 = \mu_2 = 4, \lambda_e = 5, p_1 = p_2 = 100, p_{e1} = p_{e2} = 40, h_1 = h_2 = 10, N = 2, \gamma = 0.6$) (b) Optimal production, rationing, and routing policy for a network with two heterogeneous manufacturers participating ($\lambda_1 = \lambda_2 = 2, \mu_1 = \mu_2 = 4, \lambda_e = 5, p_1 = 70, p_2 = 100, p_{e1} = p_{e2} = 35, h_1 = h_2 = 10, N = 2, \gamma = 0.6$).

hand side graph, Figure 2.4(a), shows the optimal policies for a network with two

homogeneous members. According to this figure, the optimal production policy is a base stock policy and the optimal base-stock level S_{cd}^* is equal to 10 for both producers. It is also optimal for the producers to satisfy the demand from their own customers whenever there is at least one item available in their inventories. If there are no parts available, producers do nothing (DN) and they could not sell any product (they should just produce the products in this part). The optimal rationing policy is to admit the external customers when the available inventory is greater than the optimal rationing value R_{cd}^* . In this example, the value of R_{cd}^* is equal to 3 for both producers. Similarly, the optimal base stock and rationing levels are the same for both producers since the producers are homogeneous.

The optimal routing policy for the network with homogeneous participants is sending the arriving external customer to the manufacturer whose on-hand inventory is the highest one over the optimal rationing level R_{cd}^* and rejecting them if the level of inventory for all participants is less than or equal to rationing value R_{cd}^* . This policy is called “Join the Shortest Queue (JSQ)” routing policy that is used in server farms with identical servers (Gupta et al, 2007). JSQ strives to balance the external demand across the participants, reducing the probability of one producer having several external requests while another producer does not have any external demand.

Figure 2.4(b) represents the structure of the optimal production, rationing, and routing policies in a network with two heterogeneous participating manufacturers. For a cooperative network with heterogeneous manufacturers, the optimal threshold levels for production, rationing, and routing policies are different for the participating manufacturers based on their own parameters. In this case, the optimal rationing level for the first producer is smaller than the second producer because the second one sells the products to his own customers with a higher price. Our numerical results show that the optimal routing policy for the network with heterogeneous producers does not follow any specific rule such as variants of JSQ and it completely depends on the parameters of the whole system.

2.6 Coalition Structure among N Manufacturers

The next step to analyze the general network is finding the optimal coalition structure which maximizes the total profit of the participants. The optimal coalition scheme between the producers may involve multiple cooperative networks to maximize the total profit while ensuring that each producer is better off compared to independent operation.

The coalition formation problem is a computationally challenging problem. There exist a range of algorithms and heuristic approaches for solving the problem efficiently (Rahwan et al, 2015). It is also very important to define a way of dividing the total profit of each coalition among its members in a way as to satisfy fairness and stability in order to derive the optimal coalition form. In this work, we focus on *maximizing the total benefit* of cooperation for the whole system and we do not discuss *how the total benefit is shared* among the members.

We now present a mathematical program to define the best coalition structure which maximizes the whole profit of the system under the static or dynamic routing policies. When there are N heterogeneous producers, let K be the set of all possible alternative networks including independent operation where there is no cooperation between the producers. We denote the set of all producers participating in network k , $k \in K$, by C_k . Let the number of producers in network k be $N_k = |C_k|$ and the producers in C_k be reindexed from 1 to N_k . We define the binary variables $Z_{i,k}$ indicating whether producer i is assigned to the alternative network k , where $k \in K$. Let $\Pi_{cx,k}$ be the total profit obtained from coalition k under the routing policy x where $x = s$ for the static routing and $x = d$ for the dynamic routing policy. Thus, the resulting mathematical programming is as follows:

$$\max \sum_{k \in K} \Pi_{cx,k} \quad (2.29)$$

Subject to

$$\sum_{k \in K} Z_{i,k} = 1 \quad \forall i \in N, \quad (2.30)$$

$$Z_{i,k} = Z_{i',k} \quad \forall i, i' \in C_k, \forall k \in K, \quad (2.31)$$

$$LZ_{i,k} \geq \Pi_{cx,k} \quad \forall i \in N, \forall k \in K, \quad (2.32)$$

$$\Pi_{cx,k}^* - \sum_{i \in C_k} \Pi_{nc_i}^* \geq -L(1 - Z_{i,k}) \quad \forall i \in C_k, \forall k \in K, \quad (2.33)$$

$$\Pi_{cx,k}^* \geq \Pi_{cx,k} \quad \forall k \in K, \quad (2.34)$$

$$Z_{i,k} \in \{0, 1\} \quad \forall i \in N, \forall k \in K, \quad (2.35)$$

$$\Pi_{cx,k} \geq 0 \quad \forall k \in K. \quad (2.36)$$

In this formulation, Equation (2.29) is the objective function that maximizes the total profit of the system for the optimal coalition structure. Equation (2.30) ensures that all producers should be assigned to one of the alternative potential networks or independent operation. Equation (2.31) implies that the firms i and i' should be mathematically assigned to network k if they are both participating in network k among other alternatives. Note that the profit of network k is only computed if all producers in C_k are assigned to cooperation k due to inequality (2.32), (L is a very large number). According to inequality (2.33), the profit obtained by coalition k is better off compared to independent operation. Inequality (2.34) states that the upper bound for the total profit of group k is $\Pi_{cx,k}^*$.

For a group of homogeneous manufacturers, the grand coalition consists of all producers will be the optimal coalition scheme if the optimal decision is entering into the network. On the other hand, for heterogeneous participants the optimal coalition will be different based on various parameters of the members.

2.7 Approximate Analysis of a Dynamic Network with a High Number of Homogeneous Producers

When all the producers are homogeneous, the performance of a dynamic cooperative network with high number of producers can be evaluated approximately by using the observations on the structure of the optimal production, rationing, and routing policies discussed in Section 2.5.2.

Based on our discussion in Section 2.5.2, for the network with identical participants, the optimal routing policy is defined by JSQ. It means that the producer with the highest number of finished items over the rationing level will receive the arriving external request. When a high number of identical manufacturers cooperate in the network, the inventory levels are expected to change in a very similar pattern and the probability of receiving an external customer from the network is expected to be almost the same among all the producers over the rationing value.

Based on the numerical results that confirm this intuition, we propose approximating the JSQ routing policy with a rationing-level routing policy which sends an arriving external customer to one of the producers, whose on hand inventories are above rationing threshold with equal probability. If there are no available producers at the time of the arrival, the order is lost. Therefore, we present an approximation method to evaluate the performance of the system that uses the rationing-based routing and threshold-based production and rationing policies. Once the policies are given, the resulting queuing network can be analyzed by using the Single Queue Approximation (SQA).

2.7.1 Single Queue Approximation

The main idea behind SQA is to analyze each queue in the network independently from the others and approximate the steady-state queue length probabilities by capturing the effect of other queues on each single queue through modeling the arriving process by a stochastic process with state dependent rates.

Gupta et al (2007) shows that for a $M/M/K/R$ model, where K is the number of servers and R is any stationary queue-length-dependent routing policy, SQA with the exact arrival rate yields the same steady state queue length distribution as in original $M/M/K/R$ model. We use this result to prove that we can approximate the steady-state probabilities of inventories for each producer in the network by defining the state-dependent arrival rates appropriately.

Proposition 2.1 *The cooperative network with N participant manufacturers under the time-stationary, rationing level based routing policy can be modeled as a $M/M/N/R$ queue where R is the rationing level routing policy of the network. Therefore, the single queue approximation with the exact state-dependent arrival rates leads to the same steady state queue length distribution as in the original $M/M/N/R$ model for the cooperative network model.*

The proof for this proposition is given in Appendix A provided in Chapter 6.

Therefore, in the case of cooperation, we can write the profit maximization problem for a single producer as given below. The subscription i is omitted in the formulation because the parameters are the same for all members.

$$\begin{aligned} \Pi_c^* &= \max_{S_c, R_c} \Pi_c \\ \text{subject to: } & \gamma_c \geq \gamma. \end{aligned} \quad (2.37)$$

where

$$\Pi_c = \lambda(Pr(I_c > 0))p + \lambda_a(Pr(I_c > R_c))p_e - h\bar{I}_c. \quad (2.38)$$

and

$$\gamma_c = 1 - Pr(I_{c1} \leq R_{c1}, I_{c2} \leq R_{c2}, \dots, I_{cN} \leq R_{cN}). \quad (2.39)$$

$$\gamma_c \cong 1 - [1 - Pr(I_c > R_c)]^N. \quad (2.40)$$

Note that the exact state-dependent arrival rate of external customers for a single producer in the network is defined by λ_a in Equation (2.38). Similar to the previous cases, the producers will decide to enter the cooperative network based on the expected profit rates. We define the decision of joining the cooperative network with the rationing-based routing policy (cr) or operating independently (nc) by $g_{r_i} \in \{cr, nc\}$. The optimal decision $g_{r_i}^*$ is determined as:

$$g_{r_i}^* = \begin{cases} cr \text{ (cooperation)} & \text{if } \Pi_c^* = \max\{\Pi_c^*, \Pi_{nc}^*\} \\ nc \text{ (no cooperation)} & \text{if } \Pi_{nc}^* = \max\{\Pi_c^*, \Pi_{nc}^*\} \end{cases}. \quad (2.41)$$

2.7.2 Approximate Analysis of the Cooperative Network

Based on Proposition 2.1, the single queue approximation yields a significant computational efficiency for analyzing the cooperative production network with N producers if a good approximation for state-dependent arrival rate can be determined. In this part, we will derive a closed-form approximation for the single queue arrival rate denoted with λ_a .

Let us assume that when an external customer is routed to one of the producers since its inventory level is greater than the rationing value, there are J producers which also have inventory levels greater than R_c among the other $N - 1$ members of the network. Since it is equally likely for all the available producers to receive the arriving second-class customer, the approximate long-run rate for receiving external customers λ_a is

$$\lambda_a = \lambda_e \left(\sum_{j=0}^{N-1} \frac{1}{j+1} \Pr(J=j) \right) \quad (2.42)$$

where J is a random variable that shows the number of available producers at the time of arrival. Since the steady-state probability of satisfying external demand for a producer is $\Pr(I_c > R_c)$, $\Pr(J=j)$ can be approximated by using the binomial distribution:

$$\Pr(J=j) = \binom{N-1}{j} (\Pr(I_c > R_c))^j (1 - \Pr(I_c > R_c))^{N-j-1}. \quad (2.43)$$

Equations (2.42) and Equation (2.43) yield the following closed-form approximation for λ_a :

$$\lambda_a = \lambda_e \frac{1 - (1 - \Pr(I_c > R_c))^N}{N \Pr(I_c > R_c)}. \quad (2.44)$$

The network will accept the external customers, if it can offer this group of customers the predetermined service level (γ) and we have:

$$\lambda_a = \begin{cases} \lambda_e \frac{1 - (1 - \Pr(I_c > R_c))^N}{N \Pr(I_c > R_c)} & \text{if } \gamma_c \geq \gamma \\ 0 & \text{if } \gamma_c < \gamma \end{cases}. \quad (2.45)$$

When the inter-arrival time distribution of external customers is approximated with an exponential distribution with the rate λ_a , the resulting model is the same as the make-to-stock $M/M/1$ queue with two classes of customers that was used in Section 2.3. Accordingly, Equation (2.2) yields $Pr(I_c > R_c)$ for a given value of $\xi_c = (\lambda + \lambda_a)/\mu$. Since λ_a is approximated by using the expression for a given $Pr(I_c > R_c)$ in Equation (2.44), combining these two equalities yields a non-linear equality for $Pr(I_c > R_c)$. For given values of S_c and R_c , this non-linear equation is solved first to determine $Pr(I_c > R_c)$. This solution yields the corresponding ξ_c that is used to determine $Pr(I_c > 0)$ and $\bar{I}_c$ respectively.

2.7.3 Accuracy of the Approximation Method

In order to evaluate the accuracy of single queue approximation method, we model a cooperative network with 2, 4, 6, 8, 10 members where each member uses one of the six pairs of base-stock levels (up to 15) and rationing levels up (up to 12). For the cases where $N = 2, 4$, the exact solution is determined by using the state-space model. For the cases $N = 6, 8, 10$, estimates for the exact solution are obtained by simulation where the simulation length is set to get results with the same number of significant digits as the approximation method. Comparison of the results show that the average errors for $Pr(I_c > 0)$, $Pr(I_c > R_c)$, $\bar{I}_c$, λ_a and γ_c are 0.5%, 2.7%, 3.9%, 0.02%, 0.03% respectively.

Figure 2.5 shows the percentage of error ($e\%$) between the value of obtained profit in Equation (2.17) and the resulting value from the single queue approximation in Equation (2.38) in 30 different cases. Each group of points gives the error values of profit for a specific network with different number of identical participants (N). We apply the best fit for these values of percentage error in Figure 2.5. This figure shows that the maximum error is 2.7% when there are two cooperating producers and decreases below 1% when there are more than seven producers. These results show that the single queue approximation is quite accurate and therefore can be used to evaluate the performance of the operation of a cooperative production network.

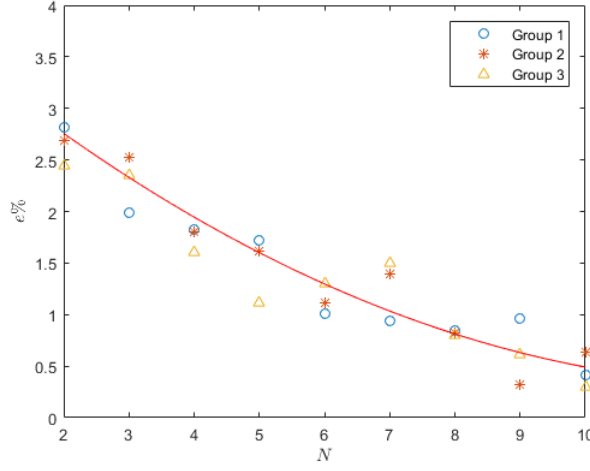


Figure 2.5: The percentage error of profit between the exact solution of dynamic network and its approximation using single queue method (Group 1: $\lambda = 3, \lambda_{ei} = 4, p = 100, p_e = 50, h = 15, \mu = 5, \gamma = 0.65$, Group 2: $\lambda = 4, \lambda_{ei} = 6, p = 100, p_e = 50, h = 15, \mu = 7, \gamma = 0.75$, Group 3: $\lambda = 5, \lambda_{ei} = 4, p = 100, p_e = 70, h = 15, \mu = 6, \gamma = 0.8$)

2.7.4 Operation of the Cooperative Production Network for a Large Number of Homogeneous Producers

The main advantage of a cooperative network is increasing the profit for all participants. By entering the network, all manufacturers receive the demand from the external customers at a higher rate. The increased external demand leads to a higher profit compared to the case where they operate independently. Our next result shows that the cooperative network with identical producers yields a profit that is greater than or equal to the profit obtained in the independent operation case.

Proposition 2.2 *The profit obtained by each producer in the cooperative network with N homogeneous participants and a rationing level routing policy is always greater than or equal to the profit obtained in the independent setting where the producers operate independently.*

The proof for this proposition is given in Appendix A provided in Chapter 6.

Given the system parameters and the production and sales policies variables, the approach described in the preceding section yields the expected profit for each producer

in both settings. When the network operates with its members optimizing their own objectives, each producer maximizes his own profit rate function to determine the optimal values of the base-stock level S_c^* and the rationing level R_c^* for given values of the price p_e and the arrival rate of the external customers λ_a given by the network. This is referred as the localized operation of the network.

The localized operation of the cooperative production network gives the same total profit as the global operation of the network where all the base stock and rationing levels are defined centrally by the network.

2.8 Benefit of Joining a Cooperative Network

In this section, we discuss the financial and service level benefits of the cooperative networks through a set of numerical experiments. To this end, we focus on a range of offered price and desired service level for the external business offer $(\lambda_{ei}, p_{ei}, \gamma)$ and discuss how the region to accept these offers is enlarged as a result of cooperation among producers.

First, we will discuss the performance of the cooperative network with the static and dynamic routing policies when the members are homogeneous. Then we will continue our discussion for the networks with heterogeneous participants and define an example for the best coalition structure among manufacturers in both static and dynamic cases. In the last part of numerical experiments, we use the approximate analytical model derived in Section 2.7 to see the benefits of cooperation among a large number of homogeneous manufacturers.

2.8.1 Comparing the Benefits of the Cooperative Networks with the Dynamic vs. Static Routing Policies and Homogeneous Members

As it was discussed in Section 2.4, the performance of a cooperative network with homogeneous producers and the static routing policy is the same as the independent operation of these producers. On the other hand, in the cooperative network with dynamic routing policy, the performance of the system will be much better especially

for the high values of the desired service level. Figure 2.6 shows the decisions to accept or reject the external business offer (λ_e, p_e, γ) for the independent operation and the cooperative network with two identical members. As it was mentioned before in Section 2.3, in the boundary graphs, the external business offer is accepted for the set of prices and requested service levels in all of the points under the boundary line. However, this offer is rejected in all combinations of prices and desired service levels over the boundary line. The right hand side graph in Figure 2.6 shows the difference in boundary lines for the static and dynamic cooperative network. As it is expected, the boundary for static network is the same as independent operation on the left hand side graph and we do not observe any improvement in system performance in this case. However, the dynamic network improves the accepting area significantly and it is very effective for accepting the external offers with high required service levels even with just two homogeneous members.

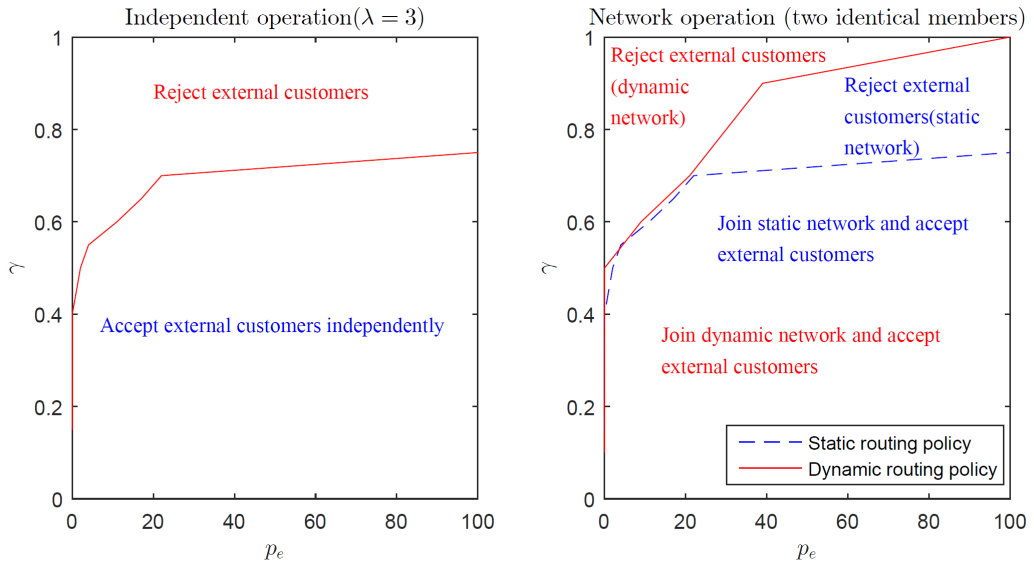


Figure 2.6: The boundary between the external customers' accept and reject regions for different values of selling price p_e and requested service level γ in independent operation, static and dynamic networks with two identical members ($\lambda_1 = \lambda_2 = 3, \mu_1 = \mu_2 = 5, \lambda_{e1} = \lambda_{e2} = 4, p_1 = p_2 = 100, h_1 = h_2 = 5$).

In a network with homogeneous members, since the inventory behavior for all

participants are similar, the cooperative network under the dynamic routing policy would not improve the system profitability significantly compared to the profit that is obtained under the static routing policy for lower values of the desired service level. On the other hand, for higher values of the desired service level, where the producers could not accept the external business offer under the static routing policy, we observe a significant improvement in the profit of dynamic network. According to Figure 2.7, for $\gamma = 0.6$, the profits for both types of network change closely in response to the different values of selling price p_e , however for $\gamma = 0.8$ (static network can not provide this service level), the whole profit of the dynamic network increases up to 80% by increasing p_e compared to static network.

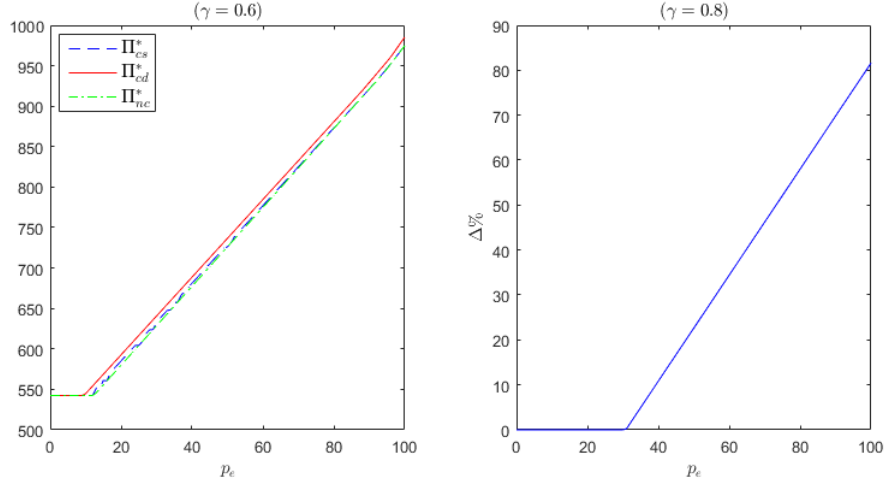


Figure 2.7: Profit rate change and its percentage of increase ($\Delta\%$) in response to different values of selling price p_e and specified (γ) in static and dynamic networks with two identical members ($\lambda_1 = \lambda_2 = 3, \mu_1 = \mu_2 = 5, \lambda_{e1} = \lambda_{e2} = 4, p_1 = p_2 = 100, h_1 = h_2 = 5$)

As a more general result, Figure 2.8 shows that the percentage increase in the profit due to using the dynamic routing policy in a homogeneous network is significantly higher for the range of the price and desired service levels that are not accepted by the network under the static routing policy.

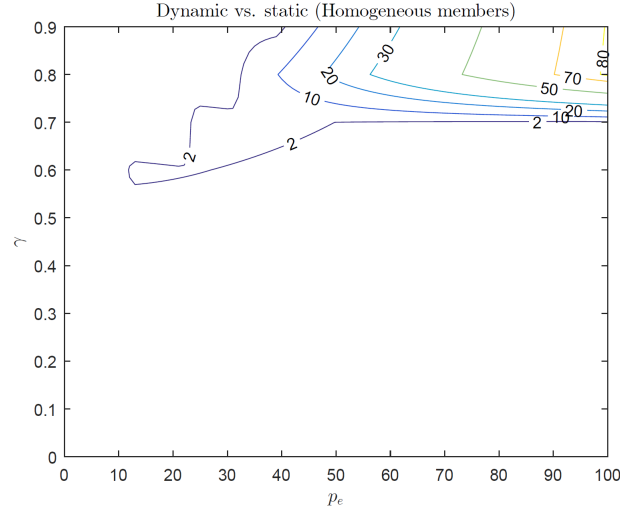


Figure 2.8: Percentage of profit increase ($\Delta\%$) in response to different values of desired service level γ and selling price p_e in a dynamic network with two identical members ($\lambda_1 = \lambda_2 = 3$, $\mu_1 = \mu_2 = 5$, $\lambda_{e1} = \lambda_{e2} = 4$, $p_1 = p_2 = 100$, $h_1 = h_2 = 5$).

2.8.2 Comparing the Benefits of Cooperative Network with the Dynamic vs. Static Routing Policies and Heterogeneous Members

In order to analyze the behavior of the cooperative network with heterogeneous members, we consider two manufacturers with different utilizations as shown in Figure 2.9. According to this figure, the region of accepting external business offers is much smaller for the second producer that receives his own customers with higher rate (higher utilization) compared to the first producer.

A cooperative network can be formed by these two producers in the areas that both manufacturers decide to join the network. Figure 2.10 shows these regions for the cooperative networks with the static and dynamic routing policies. Comparing the areas with independent operations in Figure 2.9 shows that the static network increases the acceptance area compared to the second producer but not compared to the first one. Due to the lower utilization of the first producer compared to the second one, the network with the static routing policy sends a larger share of the external customers to the first producer. Clearly, it is not optimal for Producer 1 to accept this

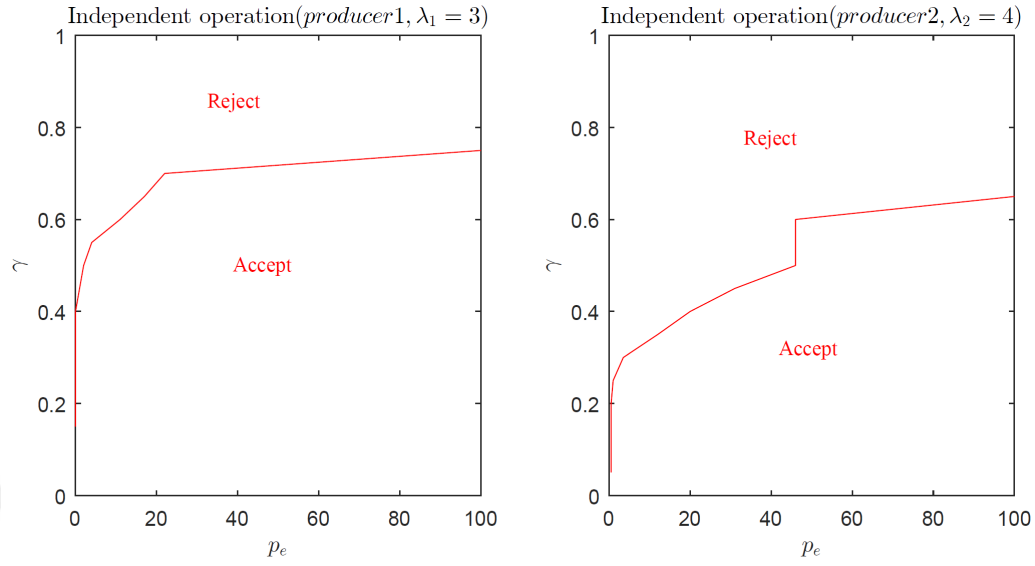


Figure 2.9: The boundary between the external offers' accept and reject regions for different values of selling price p_e and requested service level (γ) in independent operation ($\lambda_1 = 3, \lambda_2 = 4, \mu_1 = \mu_2 = 5, \lambda_{e1} = \lambda_{e2} = 4, p_1 = p_2 = 100, h_1 = h_2 = 5$).

situation whenever selling to the external customers leads to losing his own customers and decreasing the profit. Therefore, Producer 2 does not contribute to the network with a static routing policy when the required service level is high. On the other hand, the network with a dynamic routing policy is much more effective for enlarging the region where the external business offers are accepted with high values of the desired service level. Utilizing the on-line inventory information in a dynamic network leads to a higher improvement since the dynamic routing policy can react to changes in the inventory levels of heterogeneous producers effectively.

Consequently, Figure 2.11 presents the percentage profit increase for the networks with dynamic and static routing policies with respect to the profit obtained in the independent operation. Again, we observe significant profit improvement for the higher service levels in the network that uses a dynamic routing policy compared to the static network.

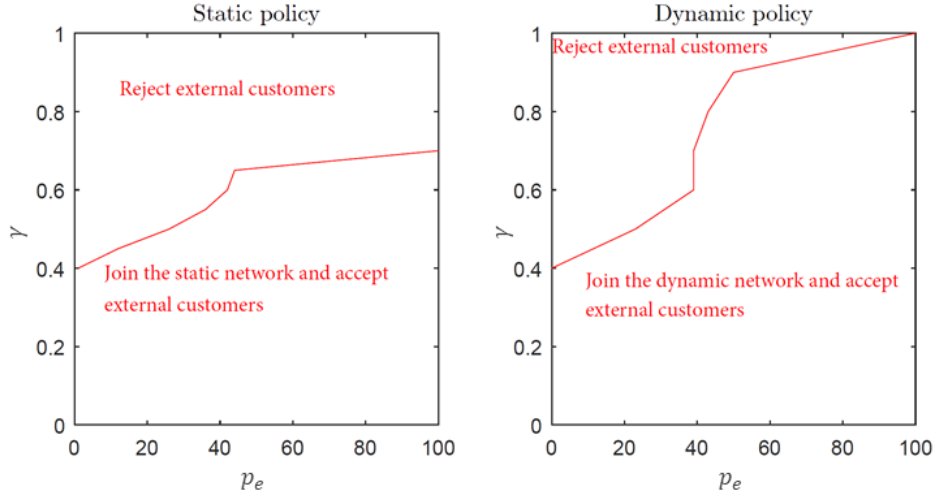


Figure 2.10: The boundary between the external offers' accept and reject regions for different values of selling price p_e and requested service level γ in static and dynamic networks with two heterogeneous members ($\mu_1 = \mu_2 = 5, \lambda_{e1} = \lambda_{e2} = 4, p_1 = p_2 = 100, h_1 = h_2 = 5$).

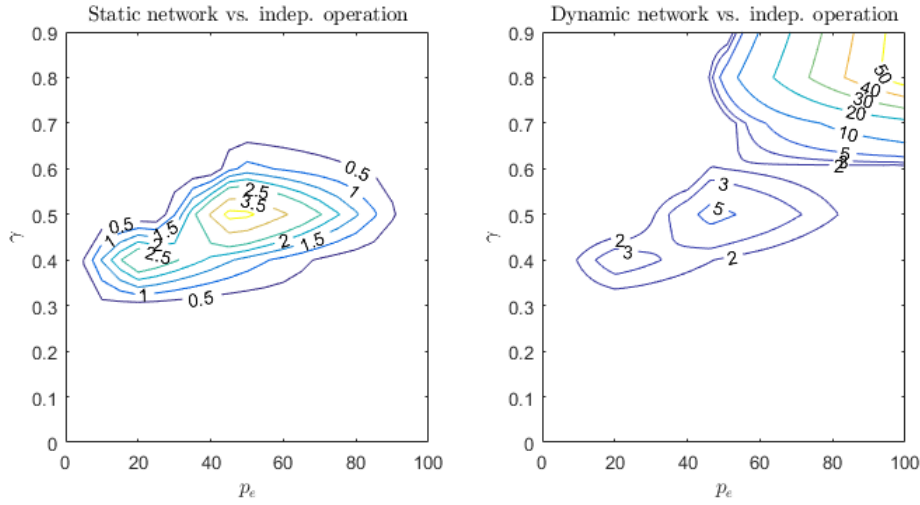


Figure 2.11: The percentage of profit increase ($\Delta\%$) for different values of selling price p_e and requested service level γ in a network with dynamic routing policy and two heterogeneous members ($\lambda_1 = 3, \lambda_2 = 4, \mu_1 = \mu_2 = 5, \lambda_{e1} = \lambda_{e2} = 4, p_1 = p_2 = 100, h_1 = h_2 = 5$).

2.8.3 Coalition Structure among Heterogeneous Manufacturers

In this section, we discuss the optimal coalition structure among heterogeneous producers. We use the mathematical programming formulation in Section 2.6 to

analyze an illustrative example with three heterogeneous manufacturers. In this example, there are 7 alternative groups C_k including the independent operations, the networks consisting only two manufacturers and the network with all three manufacturers as shown in Table 2.1. The optimal total profit is also given for all possible coalition scenarios in both the static and dynamic types of networks in Table 2.2. Based on the result of the mathematical formulation for the static case, the

Table 2.1: The optimal total profit of each alternative group C_k among three heterogeneous producers ($\lambda_1 = 2, \lambda_2 = 3, \lambda_3 = 4, \lambda_{e1} = \lambda_{e2} = 4, \lambda_{e3} = 5, \mu_1 = \mu_2 = 4, \mu_3 = 6, p_1 = 100, p_2 = 90, p_3 = 85, p_{e1} = p_{e2} = 75, p_{e3} = 70, h_1 = h_2 = h_3 = 15, \gamma = 0.6$).

Optimal total profit of each possible group C_k				
k	C_k	$\Pi_{cs,k}^*$	$\Pi_{cd,k}^*$	Π_{nc}^*
1	{1}	---	---	303.4
2	{2}	---	---	203.6
3	{3}	---	---	430
4	{2, 3}	737.1	751.8	---
5	{1, 2}	614.3	626.7	---
6	{1, 3}	738.6	757.1	---
7	{1, 2, 3}	1041.3	1076.2	---

optimal strategy for Producer 1 and 2 is to form a cooperative network and the optimal strategy for Producer 3 is to work independently. The total profit for whole system under this coalition scenario is equal to 1044.3 which dominates all other alternatives. On the other hand, for the dynamic case, the optimal strategy is to form the grand coalition with all three manufacturers. The whole profit in this case is equal to 1076.2.

Considering the system under the dynamic and static routing policies leads to different optimal coalition forms in this example. Hence it is very important to analyze the system based on the level of information exchange between the network and participants.

Table 2.2: The optimal total profit of possible coalitions among three heterogeneous producers ($\lambda_1 = 2, \lambda_2 = 3, \lambda_3 = 4, \lambda_{e1} = \lambda_{e2} = 4, \lambda_{e3} = 5, \mu_1 = \mu_2 = 4, \mu_3 = 6, p_1 = 100, p_2 = 90, p_3 = 85, p_{e1} = p_{e2} = 75, p_{e3} = 70, h_1 = h_2 = h_3 = 15, \gamma = 0.6$).

Optimal total profit for each possible coalition scenario		
Coalition structure	Optimal total profit (static)	Optimal total profit (dynamic)
$\{1\}\{2\}\{3\}$	937	937
$\{1\}\{2, 3\}$	1040.6	1055.3
$\{3\}\{1, 2\}$	1044.3	1056.8
$\{2\}\{1, 3\}$	942.2	960.7
$\{1, 2, 3\}$	1041.3	1076.2

2.8.4 Benefits of a Cooperative Network with a High Number of Homogeneous Participants

In this section, we use the approximation method presented in Section 2.7 to evaluate the benefit of a cooperative network with a high number of identical participants.

Figure 2.12 shows the decisions to accept or reject the external customers' business offer (λ_e, p_e, γ) in the cooperative network with 10 members. This figure is given for different values of external customers' arrival rate λ_e and different utilization of the participating producers (ρ) . Based on these results and focusing on the acceptance area in these two settings and corresponding independent operations in Figure 2.3, we can see that the network will accept a larger portion of the business offers from external customers with lower prices and higher service levels. As a result, the network improves the profit compared to the independent operation, this improvement is significant for the producers with higher utilization, that are not able to provide the high values of the desired service level for the external customers.

Increasing the number of participating manufacturers (N) in the cooperative network, the system will behave in different ways based on its parameters. Figure 2.13

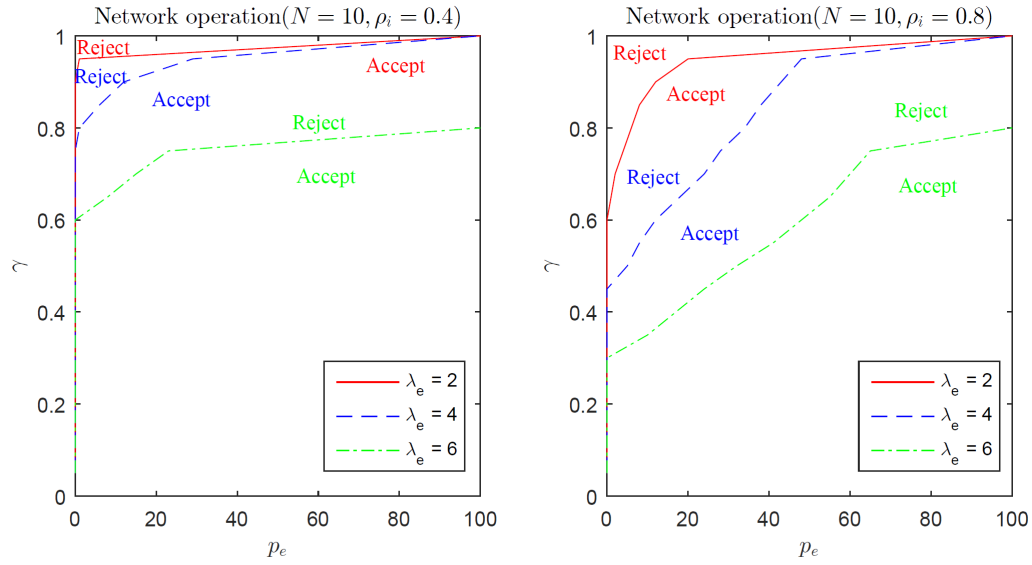


Figure 2.12: External customers acceptance and rejection areas for network based operations ($\mu = 5, p = 100, h = 5, N = 10$).

shows the optimal behavior of the system for different values of $N = 2, \dots, 100$ when $\rho_i = 0.4$ and $\lambda_{ei} = 2$. The desired service level for external customers is equal to 0.8. In this case, the producers can accept external business offer and satisfy the requested service level independently. Entering to the network gives the producers an opportunity to decrease their individually offered service level for external customers ($Pr(I_c > R_c)$), while the network offers them their desired service level $\gamma_c \geq \gamma$. According to Figure 2.13, increasing the number of participants, the producers can decrease the value of ($Pr(I_c > R_c)$) down to the constant value 0.2. As the result, the profit for the members will increase and the offered service level for external customers approaches to 1. Thus, it is more efficient to accept external business offers with high values of desired service level when there is a large number of members in the network under these circumstances.

However, it may be not true for some kinds of occasions where the producers could not satisfy external business offer independently due to service level constraint. Figure 2.14 shows the optimal behavior of the same system for different values of

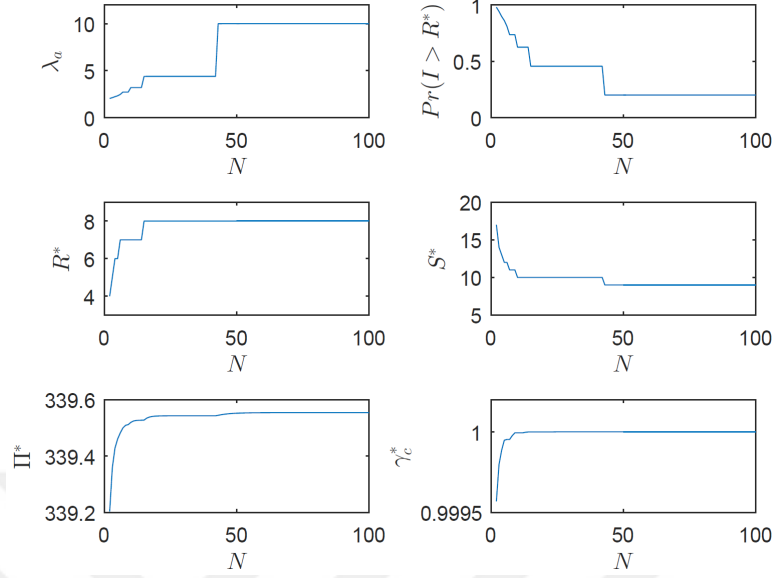


Figure 2.13: Optimal behavior of the system for different number of members in the network ($\lambda = 2, \lambda_e = 2, \mu = 5, p = 100, p_e = 70, h = 0.05, \gamma = 0.8$).

$N = 2, \dots, 100$ where $\lambda_{ei} = 6$. According to this figure, up to $N = 15$, the network is not able to provide requested service level and rejects external business offer. From $N = 16$, the service level constraint is satisfied and the network accepts external customers. By increasing the number of participating members in the network to the larger values, the value of $(Pr(I_c > R_c))$ approaches to zero in this situation. Due to this reason, the value of γ_c goes to a limited value less than 1 and the profit does not increase for larger number of participants. Therefore, the number of participants in this kind of networks should be defined as the minimum number which is essential to satisfy service level constraint.

When the price of the external customers, p_e becomes lower compared to the price obtained from the producers' own customers, the producers allocate a smaller part of their inventory for the external demand. However, having a large number of participants in the system enables the network to offer a high service level for the external customers. Figure 2.15 shows the service level benefit of a cooperative network with 10 members compared to the independent operation clearly. These graphs show

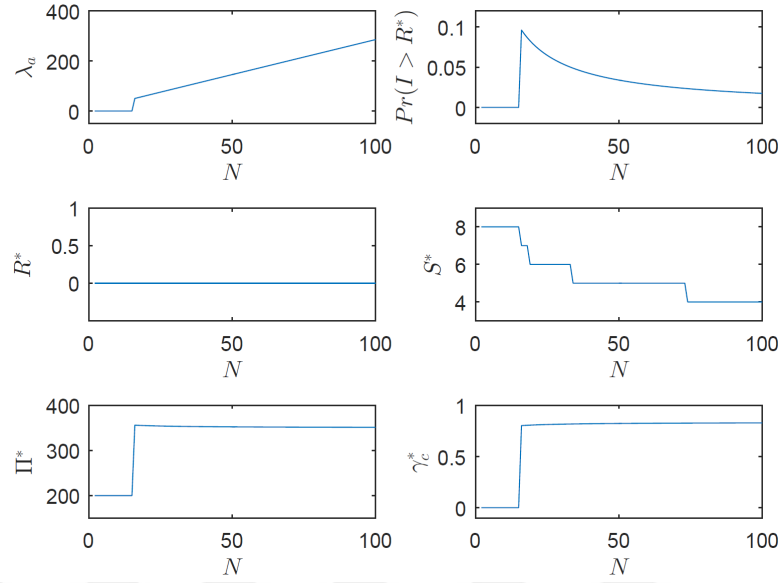


Figure 2.14: Optimal behavior of the system for different number of members in the network ($\lambda = 2, \lambda_e = 6, \mu = 5, p = 100, p_e = 70, h = 0.05, \gamma = 0.8$).

that for the same price, the network can offer a higher service level to the external customers compared to the service level that producers offer to this group of customers independently.

Figure 2.16 (a) shows that the percentage increase in the profit ($\Delta\%$) for producers participating in a cooperative network with 10 participants decreases with the price ratio p/p_e . If the price obtained from the producers' own customers is much higher than the price obtained from the external customers, the producer prefers to serve his own customers and the system gets closer to the one with a single demand class. Hence, the network does not yield more benefit. Figure 2.16 (b) shows the percentage increase in the profit ($\Delta\%$) for different values of p_e when p is equal to 100. This result also shows that entering the cooperative network is more beneficial for producers as the price for the external customers increases and gets closer to the price obtained from their own customers, p .

Figure 2.17 also shows the percentage profit increase ($\Delta\%$) for different values of price (p_e) and the requested service level (γ). According to this graph, for the

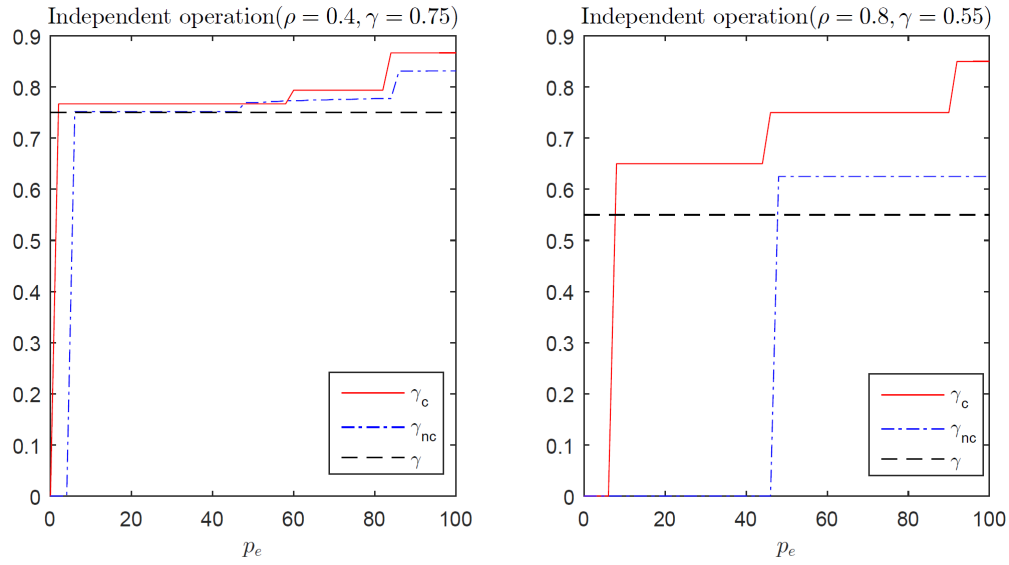


Figure 2.15: Network based service level and independent producer based service level for different values of price ($\mu = 5, \lambda_{e_i} = 4, p = 100, h = 5$).

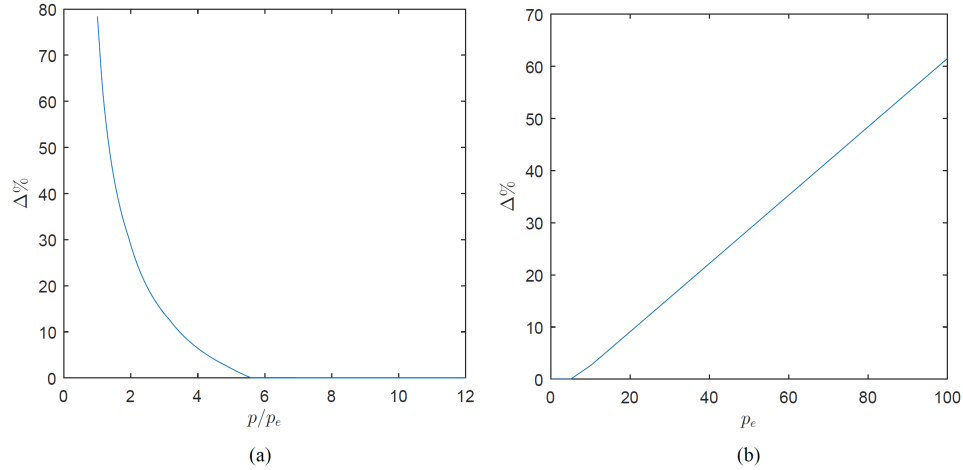


Figure 2.16: (a) Percentage of profit increase for cooperative network compared to independent operation for different values of price ratio p/p_e (b) Percentage of profit increase for cooperative network compared to independent operation for different values of price p_e ($\lambda = 4, \mu = 5, \lambda_e = 20, p = 100, h = 5, \gamma = 0.85, N = 10$).

lower levels of the requested service level, the profit increase as a result of joining the network is limited. As the requested service level increases, the independent producers

start rejecting the external business offers for the lower values of p_e . On the other hand, the network continues accepting these offers even for lower prices. This fact leads to larger percentage of profit increase for higher requested service levels.

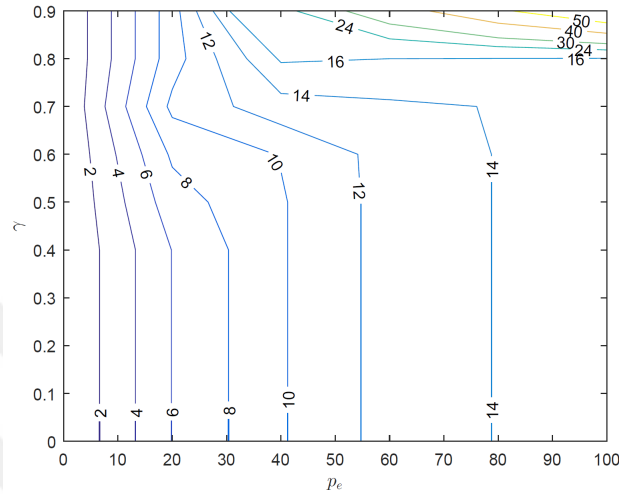


Figure 2.17: Profit increase $\Delta\%$ for different values of price p_e and requested service level γ ($\lambda = 4, \mu = 5, \lambda_{ei} = 2, p = 100, h = 5, N = 10$).

2.8.5 Concluding Remarks

In this part, we summarize our findings that may provide insights. We group our findings for cooperative networks among homogeneous producers and cooperation among heterogeneous producers.

When homogeneous manufacturers consider cooperation,

- if it is optimal for them to accept the external offers when they operate independently, it is always beneficial to join the cooperative network;
- the cooperative network will be more beneficial when the price for the external customers gets closer to the price received from the producers' own customers;
- collaboration in a network that uses a static routing policy does not bring financial and service level benefits;

- collaboration in a network that uses a dynamic routing policy with the on-line inventory information of the network members brings significant benefits especially when the requested service level is high.

When heterogeneous manufacturers with different parameters consider cooperation,

- the benefits of the static cooperative network will be highly dependent on different features of the participants;
- in order to have a successful cooperative network, each partner should contribute either by its higher external demand rate, higher production capacity or more profitable external sales;
- if a producer dominates others in all these features, cooperation between these producers can decrease the overall total profit;
- using a dynamic routing policy that utilizes the on-line inventory information in a cooperative network leads to a higher improvement for a network with heterogeneous members compared to a network with homogeneous members, since the dynamic routing policy can react to changes in the inventory levels of heterogeneous producers effectively.

Chapter 3

SIMULATION AND OPTIMIZATION OF CONTINUOUS FLOW PRODUCTION SYSTEMS BY USING MATHEMATICAL PROGRAMMING

Models that consider continuous flow of materials are used to analyze systems with both continuous and discrete materials. When the inter-arrival times are much shorter compared to other random events in the system, such as machine breakdowns, continuous flow models yield accurate representation of these systems. For example, Bierbooms (2012) discusses using continuous material flow models to analyze a beer bottling line, an assembly line for small electronic components at and a multi-server production line.

In this study, we analyze a production system with two stations separated by a finite buffer and continuous material flow. Each station has multiple modes of operation that describe various states where the stations operate with different processing rates in each state. The transition times between different states of each station are generally distributed. Since the transition times are generally distributed, an analytical solution is not available for this system. Although it is possible to simulate the system by using a general discrete event simulator, evaluating the performance and optimizing the system parameters will require extensive computational time. As an alternative, we present a mathematical programming based analytical simulation approach that captures the main dynamics of the system analytically and then find the simulated sample trajectory as a solution of a set of equations. The system parameters such as the buffer size can be optimized with respect to different objective functions by using the mathematical programming approach. The mathematical programming

representation also allows formal analysis of the properties of system performance in an alternative way. We use this representation to show that the production rate is a monotonically increasing function of the buffer size.

3.1 Past work

When the transition times between different states of each station are exponentially distributed, a continuous time, continuous-discrete state Markovian model can be analyzed analytically. There are several studies that analyze various models of an unreliable two-station continuous material flow system with a finite interstage buffer, e.g. see Wijngaard (1979), Gershwin and Berman (1981), Yeralan et al (1986), Mitra (1988), Yeralan and Tan (1997), LeBihan and Dallery (2000), Özdoğru and Altıok (2003), Helber and Jusić (2004), Diamantidis et al (2004), Tolio et al (2002), and Gershwin and Werner (2007) among others. In most of these studies, two states are used for the stations: a single up and a single down states.

There are also several studies in the literature that use multiple up and down states to model production systems in a more detailed manner by either extending the number of states in each station, incorporating phase-type distributions, or including quality dynamics into the model. Dallery (1994) studied modeling failure and repair times in stochastic models of manufacturing systems with more than two states. Özdoğru and Altıok (2003) analyzed multi state, two-valve fluid flow systems with phase-type repair times. Levantesi et al (2003) examined performance evaluation of continuous production lines with machines having different processing times and multiple failure modes. Poffe and Gershwin (2005) investigated integrating quality and quantity modeling in a production line and used the general model with multiple up and down states. Colledani and Gershwin (2013) analyzed multi-stage lines of general complex Markovian machines with finite capacity buffers. Fleuren et al (2014) presented an approximation method for fluid flow production lines with multi-server workstations and finite buffers. Kim and Gershwin (2005) analyzed single and two machine production systems based on a continuous flow model. In addition to

production systems, fluid flow models with a finite buffer are also used in computer and telecommunication systems, for example Anick et al (1982) and Elwalid and Mitra (1991).

Tan and Gershwin (2009) presented a general methodology to analyze Markovian continuous flow production systems with a finite buffer. This method uses state transition rate matrices for each station and a flow rate vector that gives the flow rate for each state as its inputs. The solution of this model gives the desired performance measures of the system. They show that various models in the literature, including the models cited above can be analyzed as special cases of this model by describing the state transition rate matrices and flow rate vectors appropriately. As a result, this representation allows analysis of a wide range of production systems (Tan and Gershwin, 2011).

The mathematical programming approach has been proposed for joint simulation and optimization of discrete event systems with continuous time and discrete state space (Schruben, 2000). Matta (2008) used a mathematical programming representation for optimization based simulation of discrete material flow production systems. Helber et al (2011) used this method to analyze and optimize stochastic flow lines. Alfieri and Matta (2012a), Alfieri and Matta (2012b) conducted two studies investigating approximate simulation of multistage production systems and pull-controlled single-product serial discrete flow manufacturing systems. Schwarz and Stolletz (2013) analyzed time-dependent stochastic flow lines using mathematical programming. Kouikoglou and Phillis (1994) presented a work about analytical simulation of unreliable serial production system with random rates.

Tan (2015) gave a mathematical programming representation of a continuous flow production system with one unreliable machine and deterministic demand. It is shown that flow rate control problems for this simple system can be solved using a mathematical programming formulation that gives a simulated sample trajectory and optimizes the parameters of the control problem for the given simulation results jointly. This study extends the approach developed by Tan (2015) to a system with

arbitrary number of states. Tan (2015) analyzed a specific system that includes one station with two states, single flow rate, and deterministic demand. Therefore, the mathematical programming formulation was developed by considering only the up and down times of a single station. In this study, we consider a general two station system where each station has multiple states of operation and different flow rates associated with each state. This extension requires a new approach for classification of system dynamics that are used as inputs and a general approach to determine the sample trajectory and the performance measures.

One way of simulating continuous flow production systems is analyzing the system in every Δt time increment within the same framework of discrete event simulation. When Δt is set to a small value, the dynamics of this discrete flow system will be close to the dynamics of a continuous flow system. However, this approach is computationally demanding. By using the sample realizations of the machines states as its input, the analytical simulation method presented in this study generates a sample realization of the buffer in an efficient way by evaluating the system at each critical time where there is a change in flow rate due to several reasons such as any change in stations' status, flow rates, or reaching full or empty buffer levels. As a result, our representational model reduces computational time significantly in comparison to discrete event representation. The mathematical programming formulation is obtained by using analytical equations that determine the critical times. The solution of this mathematical program yields simulated sample realization of this system and for the cases, where optimization of the system parameters are of interest, this approach allows joint simulation and optimization of the system.

3.2 General Model

We consider a continuous flow system with two stations, the upstream station M_u and the downstream station M_d , separated with a finite buffer with capacity Z . The state of the system at time t is represented by $s(t) = (x(t), \alpha_u(t), \alpha_d(t))$ in which $0 \leq x(t) \leq Z$ is the buffer level, $\alpha_u(t) \in \{1, \dots, I_u\}$ is the state of first station M_u and

$\alpha_d(t) \in \{1, \dots, I_d\}$ is the state of second station M_d .

In state $i \in \{1, \dots, I_u\}$, the maximum flow rate that M_u can operate with is $\mu_i^u(t)$. Similarly, in state $j \in \{1, \dots, I_d\}$, the maximum flow rate that M_d can operate with is $\mu_j^d(t)$. The actual flow rate of a station in state $(x(t), \alpha_u(t), \alpha_d(t))$ can be different depending on whether it is starved or blocked. Let the states of M_u and M_d at time t be $(\alpha_u(t), \alpha_d(t)) = (\beta, \gamma)$. When the buffer is empty with $\mu_\beta^u(t) = 0$ and $\mu_\gamma^d(t) > 0$ then M_d is forced to stop and becomes completely starved; when the buffer is full in machine state with $\mu_\beta^u(t) > 0$ and $\mu_\gamma^d(t) = 0$ then the flow into the buffer is stopped and M_u is completely blocked. However there are many cases in which the stages become partially starved or blocked but can continue with a reduced production rate. For example, when the buffer is empty and $\mu_\gamma^d(t) > \mu_\beta^u(t) > 0$, the downstream station operates at a reduced flow rate of $\mu_\beta^u(t)$. Similarly, when the buffer is full and $\mu_\beta^u(t) > \mu_\gamma^d(t) > 0$, the upstream station operates at a reduced flow rate of $\mu_\gamma^d(t)$.

The transition time to change the state of the system from $\alpha_u(t) \in \{1, \dots, I_u\}$ to $\alpha_u(t) \in \{1, \dots, I_u\}$ for M_u and the transition time to change the state of the system from $\alpha_d(t) \in \{1, \dots, I_d\}$ to $\alpha_d(t) \in \{1, \dots, I_d\}$ for M_d are random variables with given general distributions. When these transition times are exponentially distributed, Tan and Gershwin (2009) derive a set of differential equations that describe the dynamics of the system and solve these equations by using a number of boundary conditions that are determined by analyzing the empty and full buffer processes.

When the transition times are generally distributed random variables, an analytical solution is not available. For the system with generally distributed random variables, we use the given sample realizations of the upstream station M_u and the downstream station M_d as the input and derive the sample realization of the buffer level $x(t)$.

3.3 System Dynamics

In order to evaluate the system, we derive a set of linear equations to determine each critical time where the buffer level or the effective flow rate into the buffer changes. Solving this set of equations with respect to the initial condition will give the dynamics

of the structure using a mathematical formulation.

3.3.1 Inputs and State Variables

The inputs used in this approach are the time instances corresponding to the instances when the state of each station changes, and the maximum flow rates of each station at these time instances.

More specifically, we start with sample realization of stations M_u and M_d that is obtained by simulating each station independently for the run time of the simulation, τ . The sample realization gives the state of each station and the corresponding flow rates at each time instant where the state changes. Let t_i be the i th time instance when the state of M_u or M_d changes and $N + 1$ be the total number of changes in the states of M_u and M_d during the simulation period. Then the inputs used in our approach are the vector $T = \{t_i\}, i = 1, \dots, N, t_{N+1}$, and the maximum flow rates for each station at these times $\{\mu_{\alpha_u(t_i)}^u(t_i), \mu_{\alpha_d(t_i)}^d(t_i)\}, i = 1, \dots, N$.

According to the sign of the difference between the maximum flow rates of M_u and M_d at the times the state of one of the stations changes, we classify the time instances in T in three sets: T^+ for the time points where the rate of the change is positive and the buffer level increases, T^- for the time points where the rate of change is negative and the buffer level decreases, and T^0 for the time points where the rate of change is zero and the buffer level does not change. Let the rate of the change in inventory level at time t_i be $u(t_i) = \mu_{\alpha_u(t_i)}^u(t_i) - \mu_{\alpha_d(t_i)}^d(t_i)$. Then $T^+ = \{t_i | u(t_i) > 0\}$, $T^- = \{t_i | u(t_i) < 0\}$ and $T^0 = \{t_i | u(t_i) = 0\}$. The set of indices for the time instances where the buffer level increases, decreases, and stay the same, I^+ , I^- , I^0 are also defined as $I^+ = \{i | t_i \in T^+\}$, $I^- = \{i | t_i \in T^-\}$ and $I^0 = \{i | t_i \in T^0\}$.

In addition, in order to determine the amount of material produced, we define $m_u(t_i)$ and $m_d(t_i)$ as indicator variables for the upstream and downstream station respectively. The indicator variable is 1 if it has a non-zero flow rate at time t_i and 0 otherwise. By using the input vectors, we also define $MO_u = \{m_u(t_i)\}, i = 1, \dots, N$ and $MO_d = \{m_d(t_i)\}, i = 1, \dots, N$ to indicate whether each station has a non-zero flow

rate at each time instance where

$$m_u(t_i) = \begin{cases} 1 & \text{if } \mu_{\alpha_u(t_i)}^u(t_i) > 0, t_i \in T \\ 0 & \text{if } \mu_{\alpha_u(t_i)}^u(t_i) = 0, t_i \in T \end{cases} \quad (3.1)$$

and

$$m_d(t_i) = \begin{cases} 1 & \text{if } \mu_{\alpha_d(t_i)}^d(t_i) > 0, t_i \in T \\ 0 & \text{if } \mu_{\alpha_d(t_i)}^d(t_i) = 0, t_i \in T \end{cases}. \quad (3.2)$$

3.3.2 Critical Times

The buffer level at a given time instance can be determined by using the buffer level at the preceding time instance, the difference between the flow rates of the upstream and downstream stations at the preceding time instance, and the period of time since the preceding time instance. Since we know that the buffer level cannot exceed the buffer capacity and it cannot become negative, we can write

$$x(t_{i+1}) = \begin{cases} \min\{x(t_i) + u(t_i)(t_{i+1} - t_i), Z\}, & i \in I^+ \\ x(t_i), & i \in I^0 \\ \max\{x(t_i) + u(t_i)(t_{i+1} - t_i), 0\}, & i \in I^- \end{cases}. \quad (3.3)$$

However, using this formulation will give us only the value of the buffer level at each time instance t_i , and it does not provide any information about the buffer level trajectory between these points. When the buffer level increases or decreases, it may reach the full or empty buffer before the next time instance where one of the states changes. At the time instances when the buffer becomes full or empty, although the states of both stations stay the same, the net flow rates into and from the buffer change due to full or partial blocking and starvation. As a result, the buffer trajectory is linearly increasing or decreasing until this time instance and stays at the same level, at Z or at 0 , until the next change. Therefore, a full characterization of the system that includes all the critical times and the corresponding buffer levels is required to determine performance measures such as production rate, average work in process, fraction of time the buffer stays full or empty.

We define tf_i , $i \in I^+$ as the time for the buffer level to reach its capacity. If the state of one of the stations changes before tf_i , i.e., if $t_{i+1} < tf_i$, the time to reach the full buffer does not change the buffer trajectory. Otherwise, tf_i is required to determine the buffer trajectory between t_i and t_{i+1} . In this case, the buffer level increases from t_i to tf_i and reaches the full buffer level at this instance, and then stay at the full buffer until t_{i+1} .

Similarly, we define te_i , $i \in I^-$ as the time for the buffer level becomes empty. If the state of one of the stations changes before te_i , i.e., if $t_{i+1} < te_i$, the time to reach the zero buffer level does not change the buffer trajectory. Otherwise, te_i is required to determine the buffer trajectory between t_i and t_{i+1} since the buffer level decreases from t_i to te_i and becomes empty, and then stay empty until t_{i+1} .

Let $T_f = \{tf_i | tf_i < t_{i+1}\}$ and $T_e = \{te_i | te_i < t_{i+1}\}$ be the sets of time instances to reach full and empty buffer level before the next state change. Then the set of critical times that completely defines the buffer trajectory is $\Gamma = T \cup T_f \cup T_e$.

With this representation, the flow rate between two successive time instances in Γ is constant. Therefore, the buffer level at the next time instance can be determined by using the flow rate and the starting buffer level. Accordingly, starting with an initial value of the buffer level, this method yields a full characterization of the system, $s(t) = (x(t), \alpha_u(t), \alpha_d(t))$, $t \in \Gamma$.

3.3.3 Dynamics of Reaching Full Buffer

At the critical time instances where the buffer level goes up, $u(t_i) > 0$, $t_i \in T^+$, the time that the current buffer level $x(t_i)$ reaches the buffer capacity Z is determined by using the flow rate between these time instances. Since the flow rate from t_i until t_{i+1} is constant, the time it takes to reach Z from $x(t_i)$ is $\frac{Z-x(t_i)}{u(t_i)}$. So the time to reach the full buffer, tf_i is given as

$$tf_i = \frac{Z - x(t_i)}{u(t_i)} + t_i, \quad i \in I^+. \quad (3.4)$$

The buffer level may not reach the full buffer if one of the states change before this time, i.e., if $tf_i > t_{i+1}$. Therefore, the next change in the buffer level will occur either

when the buffer level reaches the buffer capacity or when the state of one of the stations changes at the next critical time instance. The length of time up to the next change is denoted with Δp_i that is given as

$$\Delta p_i = \begin{cases} t_{i+1} - t_i & \text{if } tf_i \geq t_{i+1} \\ tf_i - t_i & \text{if } tf_i < t_{i+1} \end{cases}, \quad i \in I^+, \quad (3.5)$$

and we set $\Delta p_i = 0$ for $i \in I^- \cup I^0$.

Accordingly, the length of the time interval in which the process remains at full buffer level Δf_i can be calculated for $i \in I^+$ as

$$\Delta f_i = \begin{cases} 0 & \text{if } tf_i \geq t_{i+1} \\ t_{i+1} - tf_i & \text{if } tf_i < t_{i+1} \end{cases}, \quad i \in I^+, \quad (3.6)$$

and we set $\Delta f_i = 0$ for $i \in I^-$.

3.3.4 Dynamics of Reaching Empty Buffer

At the critical time instances where the buffer level goes down, $u(t_i) < 0$, $t_i \in T^-$, the time that the buffer level $x(t_i)$ reaches 0 is also determined by using the flow rate at time t_i . Since the flow rate from t_i until t_{i+1} , $u(t_i) < 0$ is constant, the time it takes to reach 0 from $x(t_i)$ is $\frac{-x(t_i)}{u(t_i)}$. So, the time to reach the empty buffer, te_i is given as

$$te_i = \frac{-x(t_i)}{u(t_i)} + t_i, \quad i \in I^-. \quad (3.7)$$

The process will reach the zero buffer level when $te_i \leq t_{i+1}$. Consequently, the length of time up to the next change, Δe_i is defined as:

$$\Delta e_i = \begin{cases} t_{i+1} - t_i & \text{if } te_i \geq t_{i+1} \\ te_i - t_i & \text{if } te_i < t_{i+1} \end{cases}, \quad i \in I^-, \quad (3.8)$$

and we set $\Delta e_i = 0$ for $i \in I^+ \cup I^0$.

The length of the time that process stays at zero level, Δz_i is calculated for $i \in I^-$ as

$$\Delta z_i = \begin{cases} 0 & \text{if } te_i \geq t_{i+1} \\ t_{i+1} - te_i & \text{if } te_i < t_{i+1} \end{cases}, \quad i \in I^-, \quad (3.9)$$

and we set $\Delta z_i = 0$ for $i \in I^+$.

3.3.5 Dynamics of Staying at the Same Level

The rate of change in the buffer level is equal to zero when the flow rates of both stations are the same. In these cases, the buffer level at the following critical time stays the same as the level of inventory in previous critical time, i.e. $x(t_{i+1}) = x(t_i)$.

When the buffer level does not change, the length of time that the buffer level stays at zero, Δz_i for $i \in I^0$, or at the maximum level, Δf_i for $i \in I^0$ are calculated as follows:

$$\Delta z_i = \begin{cases} 0 & \text{if } x(t_i) \neq 0 \\ t_{i+1} - t_i & \text{if } x(t_i) = 0 \end{cases}, \quad i \in I^0, \quad (3.10)$$

$$\Delta f_i = \begin{cases} 0 & \text{if } x(t_i) \neq Z \\ t_{i+1} - t_i & \text{if } x(t_i) = Z \end{cases}, \quad i \in I^0. \quad (3.11)$$

Since the buffer level also stays the same when both of the stations have a flow rate of 0, an output is produced only when the flow rates are the same and greater than 0. The state variable Δh_i is also defined as the time duration of producing material when the rate of the change in inventory level is equal to zero:

$$\Delta h_i = m_u(t_i).m_d(t_i).(t_{i+1} - t_i), \quad i \in I^0, \quad (3.12)$$

and we set $\Delta h_i = 0$ for $i \in I^+, I^-$.

3.3.6 Buffer Level Dynamics

As a result, using these definitions and also the given critical times, the main buffer level dynamics are written as a set of linear equations as given below:

$$x(t_{i+1}) = \begin{cases} x(t_i) + u(t_i)\Delta p_i, & i \in I^+ \\ x(t_i), & i \in I^0 \\ x(t_i) + u(t_i)\Delta e_i, & i \in I^- \end{cases}. \quad (3.13)$$

By definition, we set $x(t f_i) = Z, i \in I^+$ and $x(t e_i) = 0, i \in I^-$.

The initial value of the buffer level is also defined as:

$$x(t_1) = x_0. \quad (3.14)$$

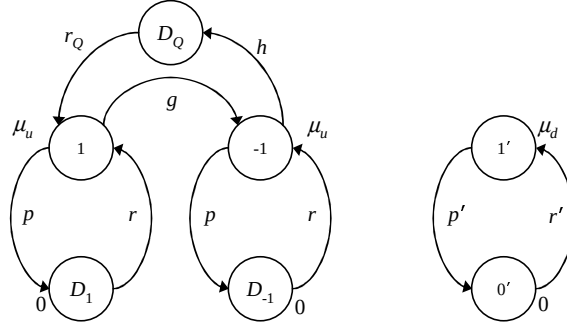


Figure 3.1: State transition diagram of the system with two stations and quality modeling

Therefore, writing these equations and solving them iteratively at each critical time t_i starting with the initial value of the buffer level will give us the trajectory of the buffer process which is determined by constructing the line segments.

3.3.7 Example: A Two Station Model with Quality Modeling

In order to explain the methodology in detail, we examine a specific model of a production system with two unreliable stations with quality specifications and a finite buffer (Poffe and Gershwin, 2005). The first station has two up (state 1 and -1) and three breakdown (D_1, D_{-1}, D_Q) modes. The processing rate of the first station for both up states is equal to μ_u and the processing rate for its down states is equal to 0. There is one up ($1'$) and one breakdown mode ($0'$) for the second station. Each of these states has the processing rate of μ_d and zero respectively. Figure 3.1 illustrates the state transitions of M_u and M_d for the proposed example.

For this specific example, there are four possible rates for the buffer level $u(t)$ between two successive critical times: $\mu_u - \mu_d$, μ_u , $-\mu_d$ and 0.

Figure 3.2 depicts a sample realization of the whole process including of stations' behavior and the resulting buffer level. The first two graphs show the state of stations on any time during the simulation period. For the sake of simplicity in this example,

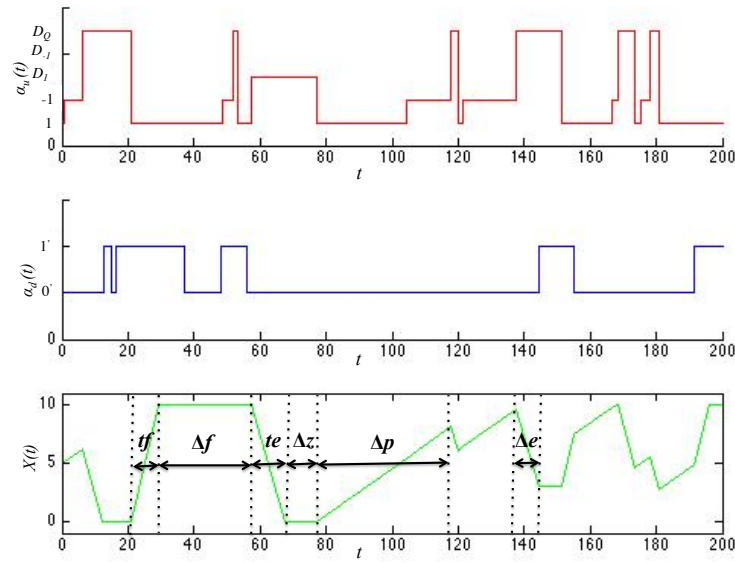


Figure 3.2: Sample realization for the system with multiple up and down states ($\mu_u = 1.2$, $\mu_d = 1$, $p = 0.01$, $r = 0.1$, $p' = 0.05$, $r' = 0.10$, $g = 0.05$, $h = 0.10$, $r_Q = 0.10$)

we use exponential distribution for the failure and repair times of the system. Different levels of buffer level and the length of the times that the buffer stays full, empty, and times to reach full and empty buffer levels are illustrated in the third graph of the same figure. This model is used in Section 3.6 to investigate the computational performance of the method we propose.

3.4 Optimization Based Simulation

In this section, we present a mathematical programming representation of the dynamics of the general two-station continuous material flow system with a finite buffer. This approach embeds the equations derived in Section 3.3 to determine the critical time points and performance measures into a mixed integer linear programming (MILP) formulation. Finding a feasible solution for the resulting MILP problem will give a simulated trajectory of the system. For a given objective function and possibly additional constraints for a selection of system parameters, the same formulation can also be used for joint simulation and optimization of the system, which will be given

in the next section.

The secondary motivation of this approach is providing an alternative representation to the system dynamics based on optimization. Instead of solving a set of equations iteratively as presented in Section 3.3, solving the same set of equations by using mathematical programming may have advantages. According to a benchmarking study, in the last 25 years (1990–2014), algorithmic advances in integer optimization combined with hardware improvements have resulted in a 200 billion factor speedup in solving mixed integer optimization problems (Bertsimas, 2014). Future improvements that will increase the computational efficiency together with the advantage of combining optimization and simulation are expected to make this approach an effective method to evaluate the performance of fluid flow systems. In addition, representing the system dynamics as a mathematical program allows us to analyze certain properties of the system performance and their dependence on the system parameters in an alternative way. However, if the only objective is to simulate the system performance, there will be no need to solve optimization problem and iterative solution in Section 3.3 will define the dynamics of system in a faster way. Therefore, the main contribution of this part is to provide the initial steps to perform joint simulation and optimization of system using mathematical programming.

3.4.1 Linear Expressions for the State Variables

In order to formulate the problem as a mixed integer linear program, the equations given in Section 3.3 should be expressed as linear equations.

The values of state variables, Δp_i , Δf_i , Δe_i and Δz_i depend on whether the buffer level reaches zero or its maximum capacity level before the next critical time or not. Therefore, they are written in a linear form by defining new non-negative variables for the differences between the time to reach full or empty buffer level and the next critical time the state of the stations changes and making it sure that these non-negative variables are not set to positive values simultaneously by using integer variables.

For the difference between tf_i and t_{i+1} we have:

$$tf_i - t_{i+1} = \Delta f_i^+ - \Delta f_i^-, \quad i \in I^+. \quad (3.15)$$

where $\Delta f_i^+ \geq 0$, $\Delta f_i^- \geq 0$ and $\Delta f_i^+ \cdot \Delta f_i^- = 0$. Hence, if $tf_i \geq t_{i+1}$ then Δf_i^+ is equal to $tf_i - t_{i+1} \geq 0$. Similarly, when $tf_i < t_{i+1}$ then Δf_i^- is equal to $t_{i+1} - tf_i > 0$. In order to ensure that Δf_i^+ and Δf_i^- are not greater than zero at the same time, the following inequalities are added:

$$\Delta f_i^+ \leq If_i L, \quad i \in I^+, \quad (3.16)$$

$$\Delta f_i^- \leq (1 - If_i) L, \quad i \in I^+. \quad (3.17)$$

Where L is a large number, and If_i is a binary indicator variable that is also a new decision variable determined by the optimization model. If_i becomes 1 when $\Delta f_i^+ \geq 0$ and 0 when $\Delta f_i^- \geq 0$. Using these equations, the state variables, Δp_i and Δf_i are specified in linear form as follows:

$$\Delta p_i = t_{i+1} - t_i - \Delta f_i^-, \quad i \in I^+, \quad (3.18)$$

$$\Delta f_i = t_{i+1} - tf_i + \Delta f_i^+, \quad i \in I^+. \quad (3.19)$$

The state variables Δe_i and Δz_i are defined in a linear form in a similar way:

$$te_i - t_{i+1} = \Delta z_i^+ - \Delta z_i^-, \quad i \in I^-. \quad (3.20)$$

The constraints for Δz_i^+ and Δz_i^- that ensure that they are not positive at the same time are written as follows:

$$\Delta z_i^+ \leq Iz_i L, \quad i \in I^-, \quad (3.21)$$

$$\Delta z_i^- \leq (1 - Iz_i) L, \quad i \in I^-. \quad (3.22)$$

where Iz_i is a binary indicator variable. Iz_i is set to 1 when $\Delta z_i^+ \geq 0$ and 0 when $\Delta z_i^- \geq 0$: For $i \in I^-$, Δe_i and Δz_i are written in linear form as follows:

$$\Delta e_i = t_{i+1} - t_i - \Delta z_i^-, \quad i \in I^-, \quad (3.23)$$

$$\Delta z_i = t_{i+1} - te_i + \Delta z_i^+, \quad i \in I^-. \quad (3.24)$$

Finally, to determine the time the buffer level stays full when the net flow rate is 0, we define an indicator variable Is_i that is set to 1 when $x(t_i) < Z$ and 0 when $x(t_i) = Z$ by using the following inequalities:

$$Z - x(t_i) \leq Is_i L, \quad i \in I^0, \quad (3.25)$$

$$L(Z - x(t_i)) \geq Is_i, \quad i \in I^0. \quad (3.26)$$

By using the indicator variable Is_i , the time the buffer stays full when the net flow rate is 0 is given as

$$\Delta f_i = (1 - Is_i)(t_{i+1} - t_i), \quad i \in I^0. \quad (3.27)$$

3.4.2 Performance Measures

When the state variables are determined and the sample trajectory of buffer level is generated by using these state variables, we can estimate the values of some of the performance measures of the system that are of interest.

The average inventory level $\bar{X}$ is obtained just by calculating the area under the sample trajectory of the buffer level, and it is given as

$$\bar{X} = \sum_{i=1}^N \frac{[(x(t_i) + x(t_{i+1}))(\Delta p_i + \Delta e_i)] + 2Z\Delta f_i}{2\tau}. \quad (3.28)$$

The estimates for the probabilities of having empty buffer, F_e and full buffer as F_z , are determined by the fraction of time duration in which the buffer level stays empty or full over run time of simulation, respectively:

$$F_e = \sum_{i \in I^-, I^0} \frac{\Delta z_i}{\tau}, \quad (3.29)$$

$$F_z = \sum_{i \in I^+, I^0} \frac{\Delta f_i}{\tau}. \quad (3.30)$$

An estimate of the production rate is obtained by determining the amount of material produced by one of the machines in the simulation run time. When the

buffer level increases, the upstream station produces with its maximum production rate until the time when the buffer level reaches Z . When the buffer level is at this level, the flow rate of the upstream station is either zero due to complete blocking that happens if the downstream station is not operational during $[t_i, t_{i+1})$ or $m_d(t_i) = 0$, or it is decreased to the flow rate of the downstream station if the downstream station is operational during $[t_i, t_{i+1})$ or $m_d(t_i) = 1$. As a result, for $t_i \in T^+$, the amount of material produced by first station, P_u^+ during the time period $[0, \tau]$ is equal to:

$$P_u^+ = \sum_{i \in I^+} (\mu_{\alpha_u(t_i)}^u(t_i) \cdot \Delta p_i + m_d(t_i) \cdot \mu_{\alpha_d(t_i)}^d(t_i) \cdot \Delta f_i^-). \quad (3.31)$$

Similarly, for $t_i \in T^-$, the produced amount of material for first station (P_u^-) is equal to

$$P_u^- = \sum_{i \in I^-} m_u(t_i) \cdot (\mu_{\alpha_u(t_i)}^u(t_i) \cdot \Delta e_i + \mu_{\alpha_u(t_i)}^u(t_i) \cdot \Delta z_i^-). \quad (3.32)$$

In the states where the flow rates of upstream and downstream station are the same, $i \in I^0$, stations continue their production if their flow rates are not zero. The amount of material produced by the upstream station when the flow rates are equal, P_u^0 is then written as

$$P_u^0 = \sum_{i \in I^0} (\Delta h_i \cdot \mu_{\alpha_u(t_i)}^u(t_i)) = \sum_{i \in I^0} (t_{i+1} - t_i) m_u(t_i) \cdot m_d(t_i) \cdot \mu_{\alpha_u(t_i)}^u(t_i). \quad (3.33)$$

Hence, the estimated production rate PR is determined as follows:

$$PR = \frac{P_u^+ + P_u^- + P_u^0}{\tau}. \quad (3.34)$$

The production rate can also be estimated by using the material produced by the downstream station in a similar way. In the long run, the amount of material produced by M_u and M_d will be equal to each other.

The mathematical programming representation can be used to analyze formal properties of the performance measures and their dependence on the system parameters. One of the well-known properties in production systems is the monotonicity of the production rate with respect to the buffer capacity. Namely, as the buffer capacity

increases, the production rate also increases. In the following proposition, we use the mathematical programming representation to show this property.

Proposition 3.1 *The production rate monotonically increases with the buffer size, i.e., if $Z \geq Z'$ then $PR \geq PR'$ where PR' is the production rate of the system where the buffer capacity is Z' .*

The proof for this proposition is given in Appendix B provided in Chapter 6.

By using a similar approach, it can also be shown that the average inventory level also increases monotonically with the increasing buffer capacity.

3.4.3 Structure of the Mathematical Program for Simulation

A sample path of the system can be determined by solving the set of Equations (3.4)-(3.14) iteratively starting with an initial buffer level x_0 for all $t_i \in T$.

Alternatively, the sample path can also be determined by finding a feasible solution to the following set of linear equations with continuous and integer variables:

$$x(t_1) = x_0 \quad (3.35)$$

$$x(t_{i+1}) = x(t_i) + \Delta p_i(u_i) \quad i \in I^+ \quad (3.36)$$

$$x(t_{i+1}) = x(t_i) + \Delta e_i(u_i) \quad i \in I^- \quad (3.37)$$

$$x(t_{i+1}) = x(t_i) \quad i \in I^0 \quad (3.38)$$

$$\Delta p_i = t_{i+1} - t_i - \Delta f_i^- \quad i \in I^+ \quad (3.39)$$

$$\Delta f_i = t_{i+1} - \frac{Z - x(t_i)}{u_i} - t_i + \Delta f_i^+ \quad i \in I^+ \quad (3.40)$$

$$\frac{Z - x(t_i)}{u_i} + t_i - t_{i+1} = \Delta f_i^+ - \Delta f_i^- \quad i \in I^+ \quad (3.41)$$

$$\Delta e_i = t_{i+1} - t_i - \Delta z_i^- \quad i \in I^- \quad (3.42)$$

$$\Delta z_i = t_{i+1} + \frac{x(t_i)}{u_i} - t_i + \Delta z_i^+ \quad i \in I^- \quad (3.43)$$

$$\frac{-x(t_i)}{u_i} + t_i - t_{i+1} = \Delta z_i^+ - \Delta z_i^- \quad i \in I^- \quad (3.44)$$

$$\Delta f_i^+ \leq L \cdot I f_i, \quad \Delta f_i^- \leq L(1 - I f_i), \quad I f_i \in \{0, 1\} \quad i \in I^+ \quad (3.45)$$

$$\Delta z_i^+ \leq L.Iz_i, \quad \Delta z_i^- \leq L(1 - Iz_i), \quad Iz_i \in \{0, 1\} \quad i \in I^- \quad (3.46)$$

$$x(t_i) \geq 0 \quad i \in I^-, I^+, I^0 \quad (3.47)$$

$$\Delta f_i^+ \geq 0, \quad \Delta f_i^- \geq 0 \quad i \in I^+ \quad (3.48)$$

$$\Delta z_i^+ \geq 0, \quad \Delta z_i^- \geq 0 \quad i \in I^- \quad (3.49)$$

$$\Delta f_i^+ = 0, \quad \Delta f_i^- = 0 \quad i \in I^0, I^- \quad (3.50)$$

$$\Delta z_i^+ = 0, \quad \Delta z_i^- = 0 \quad i \in I^0, I^+ \quad (3.51)$$

When the problem is solved by using optimization, any objective function can be used for the mathematical program to determine the feasible solution. Since either Δf_i^+ or Δf_i^- , similarly either Δz_i^+ or Δz_i^- will be set to zero in the final solution, our computational experiments show that using an objective function that moves the state variables Δf_i^+ , Δf_i^- , Δz_i^+ and Δz_i^- to be close to zero provides computational advantages. Accordingly, we suggest the following objective function and solve this minimization problem for determining a sample trajectory of the system in our numerical experiments:

ProbST:

$$\text{Minimize } \sum_{i=1}^N (\Delta f_i^+ + \Delta f_i^- + \Delta z_i^+ + \Delta z_i^-) \quad (3.52)$$

Subject to

$$\text{Equations (3.35) – (3.51)}$$

As a result, the optimization problem for determining the operation of system is a mixed-integer linear program with a linear objective function, linear equality, inequality and non-negativity constraints.

By solving this optimization problem we obtain the buffer levels for all of the state variables in $i \in I^+, I^0, I^-$. The vectors of the critical times T_f and T_e that affect the trajectory are constructed by using non-zero values of $\Delta f_i, i \in I^+$ and Δz_i in the solution. That is

$$t_{f_{i-1}} = t_i - \Delta f_i, i \in I^+, \Delta f_i > 0, \quad (3.53)$$

$$te_{j-1} = t_j - \Delta z_j, j \in I^-, \Delta z_j > 0. \quad (3.54)$$

The corresponding buffer levels are also added to the trajectory automatically as $x(tf_{i-1}) = Z$ and $x(te_{j-1}) = 0$. The full characterization of the system is determined by the values of the variables $x(t_i), t_i \in T$, $x(tf_i), tf_i \in T_f$, $x(te_i), te_i \in T_e$. The performance measures can be determined directly from the solution.

The next proposition shows that the solution of the mathematical program given above is the unique solution of the set of equations that describe the systems dynamics given in Section 3.3:

Proposition 3.2 *The solution of the problem **ProbST** is unique and identical to the solution of the set of equations (3.4)-(3.14) that describe the system dynamics.*

The proof for this proposition is given in Appendix B provided in Chapter 6.

3.4.4 The Structure of Mathematical Programming Problem

The general model of optimization problem to simulate the dynamics of system is a mixed-integer linear program with a linear objective function, linear equality, inequality and non-negativity constraints. The number of continuous and integer decision variables and the number of equality and inequality constraints depend on the number of elements in the sets of critical times T^+ , T^- and T^0 generated during the run duration τ . Table 3.1 gives the number of decision variables and constraints for the problem **ProbST**. Table 3.1 shows that the size of the problem to be solved

Table 3.1: Problem Structure Based on the Number of Elements in T^+ , T^- , and T^0

Number of continuous decision variables	$4 T^+ + 4 T^- + T^0 $
Number of discrete decision variables	$ T^+ + T^- $
Number of equality constraints	$3 T^+ + 3 T^- + T^0 + 1$
Number of inequality constraints	$5 T^+ + 5 T^- + T^0 $

only depends on the number of times the buffer level changes during the run duration.

Let n_{ji}^u and n_{ji}^d be the number of total changes from state j to state i in the upstream and downstream stations respectively. Then the total number of critical times for the system can be defined as follows:

$$|T^+| + |T^-| + |T^0| = \sum_{i \in \alpha_u(t)} \sum_{j \in \alpha_u(t)} n_{ji}^u + \sum_{i \in \alpha_d(t)} \sum_{j \in \alpha_d(t)} n_{ji}^d.$$

In a Markovian system, we can determine the number of changes that will be observed analytically. Let the transition rate from state i to state j be λ_{ij} . By following an ergodicity argument, the steady-state probability of state i , denoted by π_i , is the fraction of time spent in state i in the long run. The expected total time spent in a given state can be determined by the total number of visits to that state and the expected sojourn time at each visit. The expected sojourn time in state i of M_u is $1 / \sum_{j \in \alpha_u(t)} \lambda_{ij}$ and the expected sojourn time in state i of M_d is $1 / \sum_{j \in \alpha_d(t)} \lambda_{ij}$. Accordingly, the steady-state probability is state $i \in \{1, \dots, I_u\}$ can be expressed as

$$\pi_i = \lim_{\tau \rightarrow \infty} \frac{\sum_{j \in \alpha_u(t)} n_{ji}^u / \sum_{j \in \alpha_u(t)} \lambda_{ij}}{\tau}, \quad i \in \{1, \dots, I_u\}.$$

Then, for a large value of τ ,

$$\tau \pi_i \sum_{j \in \alpha_u(t)} \lambda_{ij} = \sum_{j \in \alpha_u(t)} n_{ji}^u.$$

By following the same steps for n_{ji}^d , we can express the total number of changes observed during the run duration as

$$|T^+| + |T^-| + |T^0| = \tau \left(\sum_{i \in \alpha_u(t)} \pi_i \sum_{j \in \alpha_u(t)} \lambda_{ij} + \sum_{i \in \alpha_d(t)} \pi_i \sum_{j \in \alpha_d(t)} \lambda_{ij} \right). \quad (3.55)$$

Equation (3.55) shows that the number of equations and constraints of the mathematical program does not depend on the structure of the system such as the number of states in the state space or the buffer capacity. It depends on the rates of change for transitions among the states of the system. In other words, it is possible to have a

mathematical program with significantly higher number of constraints and variables for a system with a few number of states for each station and a small buffer if the rates of change among the states are high. Conversely, the mathematical program of a system with large number of states for each stage and a large buffer may have much fewer variables and constraints if the rates of changes are not that high. The relationship between the number of variables and number of constraints and the computation time is investigated numerically in Section 3.6.

In order to determine the design parameters according to a given objective and possibly with additional constraints, the corresponding objective function replaces the one given above and the additional constraints are added to the set. Two design problems are discussed in the next section.

3.5 Joint Simulation and Optimization

In this section, we use the mathematical programming approach to perform joint simulation and optimization of the system. The solution of the optimization problems provides the simulated trajectory and also the optimal values for the objective function.

In Section 3.4, the optimization problem was formulated for a given buffer capacity level Z . In this section, we consider the buffer capacity determination problem where the buffer capacity Z is added as an additional decision variable for the problem, and its optimal value is determined based on two different objective functions. In the first one, the minimum buffer capacity is set to achieve a predetermined value of the production rate. In the second one, the optimal buffer capacity is determined to achieve maximum profit that is the difference between the revenue generated from the production and the inventory cost.

3.5.1 Minimum Buffer Capacity Determination Problem to Achieve the Desired Production Rate

When the buffer capacity is a decision variable, and the objective function is minimizing the buffer capacity to achieve a desired production rate PR_{given} , the final formulation

includes the same set of constraints as the simulation mathematical program given in Section 3.4.3 in addition to additional constraints related to production rate calculation for a given value of the buffer capacity. **ProbMinBufferPR:**

$$\text{Minimize } Z \tag{3.56}$$

Subject to

Equations (3.35) – (3.51)

$$\Delta p_i = 0, \Delta f_i^- = 0 \quad i \in I^-, I^0 \tag{3.57}$$

$$\Delta e_i = 0, \Delta z_i^- = 0 \quad i \in I^+, I^0 \tag{3.58}$$

$$\Delta h_i = m_u(t_i) \cdot m_d(t_i) \cdot (t_{i+1} - t_i) \quad i \in I^0 \tag{3.59}$$

$$\Delta h_i = 0 \quad i \in I^+, I^- \tag{3.60}$$

$$P_u^+ = \sum_{i=1}^N (\mu_{\alpha_u(t_i)}^u(t_i) \cdot \Delta p_i + m_d(t_i) \cdot \mu_{\alpha_d(t_i)}^d(t_i) \cdot \Delta f_i^-) \tag{3.61}$$

$$P_u^- = \sum_{i=1}^N m_u(t_i) \cdot (\mu_{\alpha_u(t_i)}^u(t_i) \cdot \Delta e_i + \mu_{\alpha_u(t_i)}^u(t_i) \cdot \Delta z_i^-) \tag{3.62}$$

$$P_u^0 = \sum_{i=1}^N (\Delta h_i \cdot \mu_{\alpha_u(t_i)}^u(t_i)) \tag{3.63}$$

$$PR = \frac{(P_u^+ + P_u^- + P_u^0)}{\tau} \tag{3.64}$$

$$PR \geq PR_{given} \tag{3.65}$$

Similar to Section 3.4, we have a mixed-integer linear program with linear equality and inequality constraints, continuous and binary variables. The solution of this problem provides us with the values of state variables simulating the performance of the system together with the optimal value of the buffer capacity level Z^* depending on the problem.

According to Proposition 3.1, production rate is monotonically increasing with respect to increasing buffer capacity level as also shown in Figure 3.7. Therefore, for a given production rate, there is a unique minimum value of buffer capacity level (Z)

that satisfies the constraint $PR \geq PR_{given}$. For a given Z , Proposition 3.2 shows that the solution of the mathematical program is also unique. As a result, we can conclude that the solution of the problem **ProbMinBufferPR** gives the unique solution of the buffer minimization problem that gives the desired production rate.

3.5.2 Optimal Buffer Capacity Determination Problem to Maximize the Average Profit

In this profit maximization problem, the optimal buffer capacity is determined to maximize the average profit. Let c_p be the profit obtained by selling a unit volume of production and c_h be the holding cost per unit of volume per time. Then the profit per unit time is $c_p PR - c_h \bar{X}$. This formulation captures the trade-off between increasing the buffer level that yields an increase in the production rate but also an increase of the average inventory level. The formulation for this optimization problem is written as follows:

ProbMaxProfit:

$$\text{Maximize } c_p PR - c_h \sum_{i=1}^N \frac{[(x(t_i) + x(t_{i+1}))(\Delta p_i + \Delta e_i)] + 2Z\Delta f_i}{2\tau} \quad (3.66)$$

Subject to:

$$\text{Equations (3.35) – (3.51)}$$

$$\text{Equations (3.57) – (3.64)}$$

$$\Delta f_i = 0 \quad i \in I^- \quad (3.67)$$

$$\Delta f_i = (1 - Is_i)(t_{i+1} - t_i) \quad i \in I^0 \quad (3.68)$$

$$Z - x(t_i) \leq L.Is_i, \quad L(Z - x(t_i)) \geq Is_i, \quad Is_i \in \{0, 1\} \quad i \in I^0 \quad (3.69)$$

Since $\bar{X}$ is a quadratic function of the state variables, the above problem is a quadratic mixed integer program with linear constraints.

In this problem, the objective function is the difference between two monotonically increasing functions of the buffer capacity. Accordingly, there is a unique value of Z

that maximizes the average profit. For each given value of Z , Proposition 3.2 shows that the solution of the mathematical program gives a unique solution. As a result, the solution of the maximization problem **ProbMaxProfit** gives the unique solution for the buffer capacity that maximizes the average profit.

3.6 Numerical Results

In this section, we use the specific model described in Section 3.3.7 to analyze the proposed method numerically. We first validate the results obtained by the proposed method by comparing the results for this model with exponential transition times with the available analytical results. In order to show the applicability of the proposed method for general models, we evaluate the performance of the same system with non-exponential transition times. Subsequently, the results of buffer capacity determination problem are investigated. Finally, we evaluate the computational performance of the proposed method.

All the numerical results are obtained by using MATLAB R2012a and GAMS win64 (when optimization is used) on a computer with 2,7 GHz Intel core i5 processor with 8 GB RAM.

3.6.1 Validation and Effects of System Parameters

For validation, we use the model described in Section 3.3.7 with exponential transition times that was developed and analyzed by (Poffe and Gershwin, 2005). Tan and Gershwin (2009) also use this system to explain their methodology and validate the general model to analyze Markovian continuous flow systems with a finite buffer.

The percentage error between the simulation results for the production rate obtained by using the methodology presented in this study and the analytical production rate obtained from Tan and Gershwin (2009) is approximately equal to 10^{-5} for $\tau = 500000$ sec.

Figure 3.3 shows the effect of increasing the processing rate of the upstream station on production rate and average inventory level. As the figure shows, the production

rate of the system will increase up to its limiting point by increasing the processing rate of station M_u . The average inventory level also increases up to the buffer capacity value as the processing rate of the upstream station increases. Figure 3.4 shows the

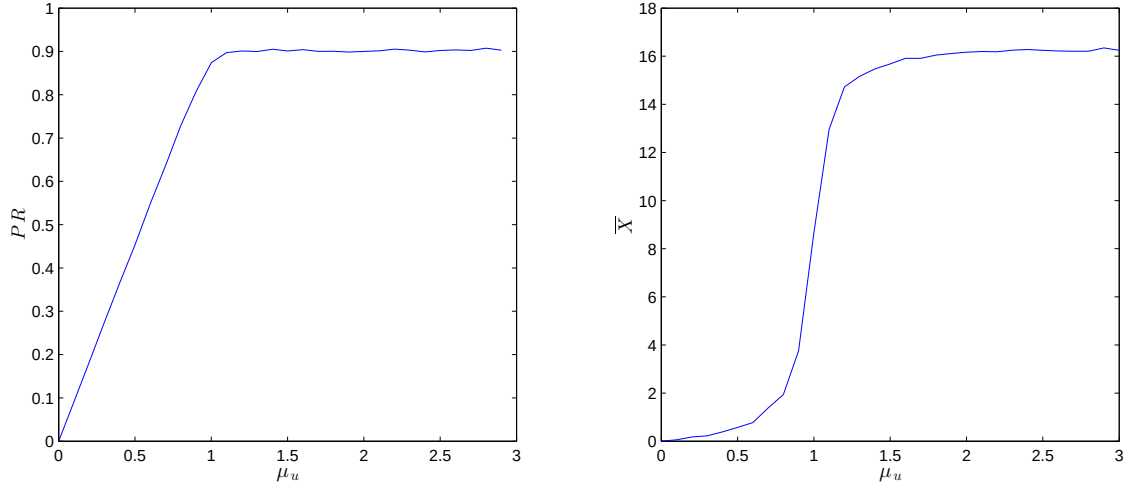


Figure 3.3: Effect of first stage processing rate on production rate and expected inventory level in Markovian model with multiple up and down states ($\mu_d = 1$, $p = 0.005$, $r = 0.15$, $p' = 0.015$, $r' = 0.15$, $g = 0.01$, $h = 0.2$, $r_Q = 0.15$, $Z = 17$)

effect of increasing the processing rate of the downstream station on production rate and average inventory level. As the figure shows, the production rate of the system will increase up to its limiting point by increasing the processing rate of station M_u . On the other hand, we can see that as the processing rate of the downstream station increases, the average inventory level decreases.

3.6.2 Results for a System with Non-exponential Transition Times

The analytical results are available only for exponential transition times as analyzed. In order to analyze a system with non-exponential transition time distributions, we use Chi-square distribution for the operational and failure times for the same model described in Section 3.3.7. In this system we have the same set of operational and failure modes as before, but, the coefficient of variation for all operational times (for the times from state 1 to D_1 , from -1 to D_{-1} , from 1 to -1, from -1 to D_Q , and from 1'

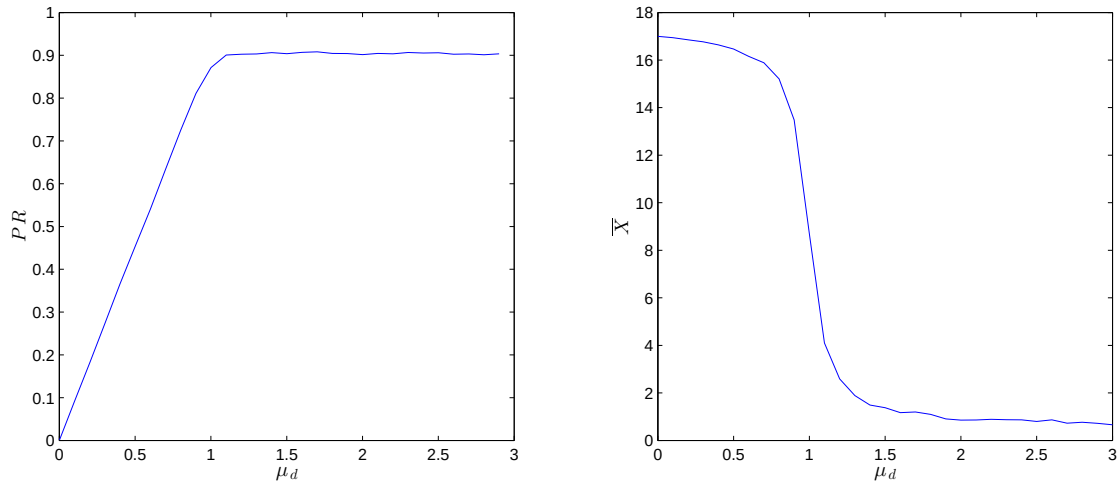


Figure 3.4: Effect of second stage processing rate on production rate and expected inventory level in Markovian model with multiple up and down states ($\mu_u = 1$, $p = 0.005$, $r = 0.15$, $p' = 0.015$, $r' = 0.15$, $g = 0.01$, $h = 0.2$, $r_Q = 0.15$, $Z = 17$)

to $0'$) are set to cv_o and the coefficient of variation for all failure times (for the times from state D_Q to 1, from D_1 to 1, from D_{-1} to -1, and from $0'$ to $1'$) are set to cv_f .

Figure 3.5 shows that as the coefficient of variation for operational times increases, the average level of inventory decreases. However, as the coefficient of variation of the failure times increases, average inventory level also increases.

Figure 3.6 shows that the production rate increases with increasing the coefficient of variation for failure times and decreases by increasing the coefficient of variation for operational times.

3.6.3 Results for Buffer Determination Problems

Figure 3.7 shows the minimum buffer capacity required to achieve a specific production rate that is determined by solving the optimization problem **ProbMinBufferPR** given in Section 3.5.1 for different desired production rate values. The optimal buffer capacity increases as the desired production rate increases up to the point where the maximum possible production rate for the system is reached. After this point increasing the buffer capacity level cannot increase the production rate. Figure 3.8

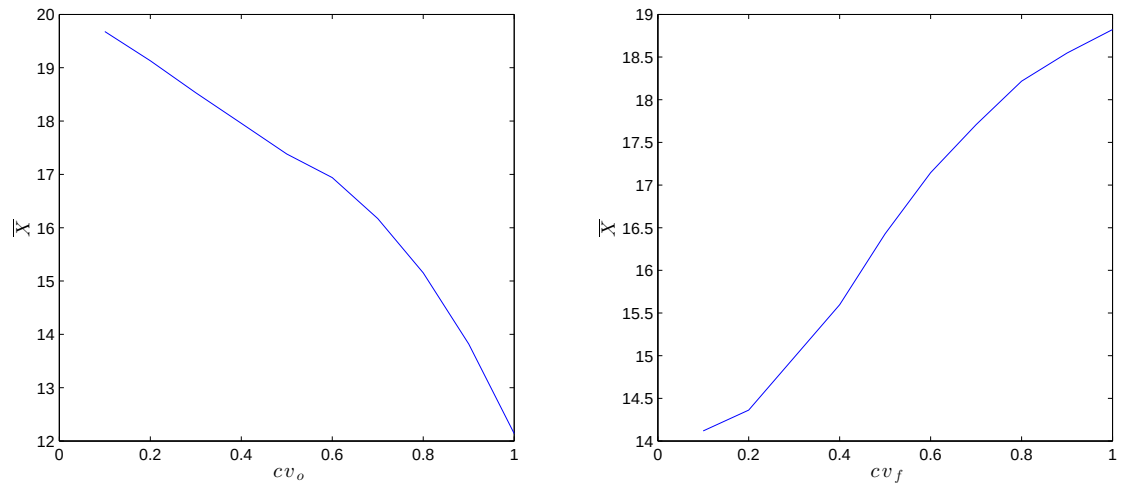


Figure 3.5: Effect of coefficient of variation on expected inventory level in general model with Chi-square distribution for operational and failure times ($\mu_u = 1.5$, $\mu_d = 1$, $Z = 20$)

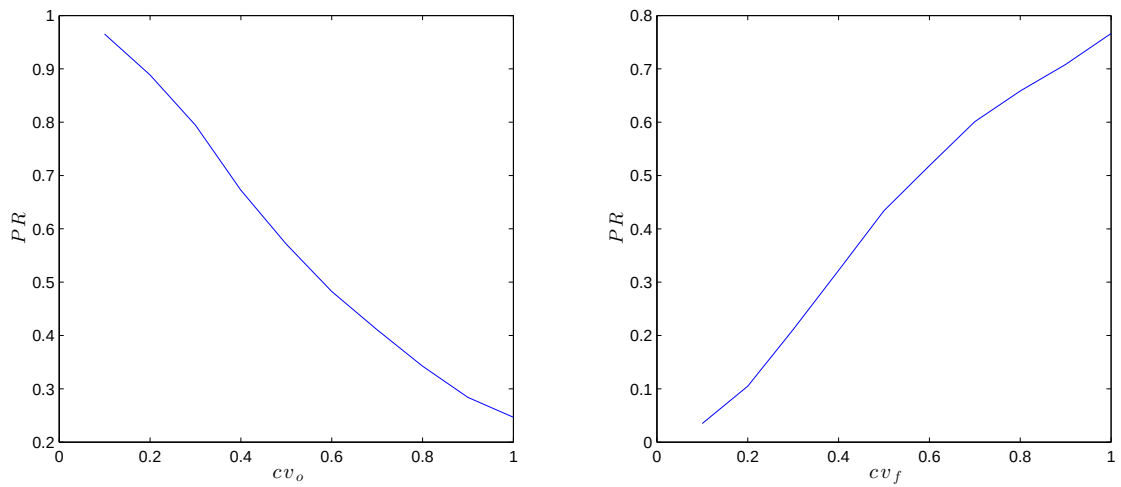


Figure 3.6: Effect of coefficient of variation on production rate in general model with Chi-square distribution for operational and failure times ($\mu_u = 1.5$, $\mu_d = 1$, $Z = 20$)

gives the results of the optimization problem **ProbMaxProfit** described in Section 3.5.2. According to these results, the production rate is optimized according to the profit obtained from the production rate and the cost of holding higher inventory to achieve a higher production rate. Increasing buffer capacity level, as decision variable

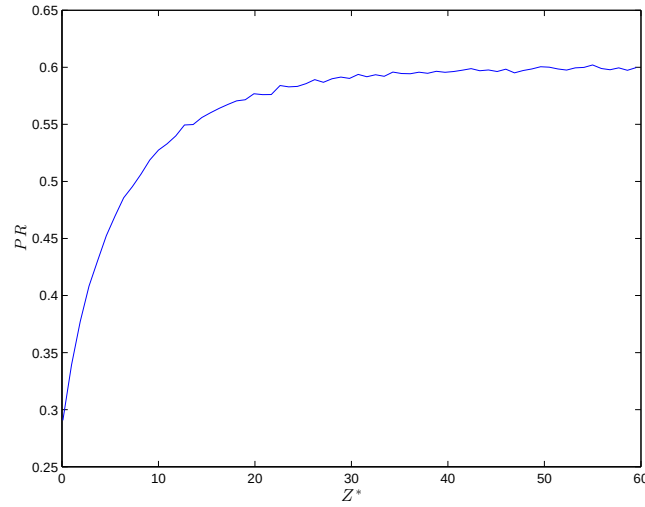


Figure 3.7: Optimal values of buffer capacity for specified production rates in the model with multiple up and down states and exponential transition times ($\mu_u = 2$, $\mu_d = 1$, $p = 0.1$, $r = 0.15$, $p' = 0.1$, $r' = 0.15$, $g = 0.1$, $h = 0.2$, $r_Q = 0.15$)

will result in higher production rate and higher level of inventory in the system. Therefore, when the holding cost c_h increases, the solution of the optimization problem achieves a balance between these increasing production rate and increasing average inventory in a way to obtain maximum profit. Since higher average inventory means higher average cycle time, there is also a trade-off between increasing the production rate and causing a longer average cycle time. According to Figure 3.8, increasing the holding cost yields a lower production rate and a shorter cycle time due to decreasing the average inventory level.

3.6.4 Computational Efficiency of Mathematical Programming Approach

Convergence:

Figure 3.9 shows the convergence of our proposed method by comparing simulation with the analytical result for the system described in Section 3.3.7 with exponential transition times. According to this figure as the length of run duration τ increases, the simulated performance converges rapidly to the exact value of production rate

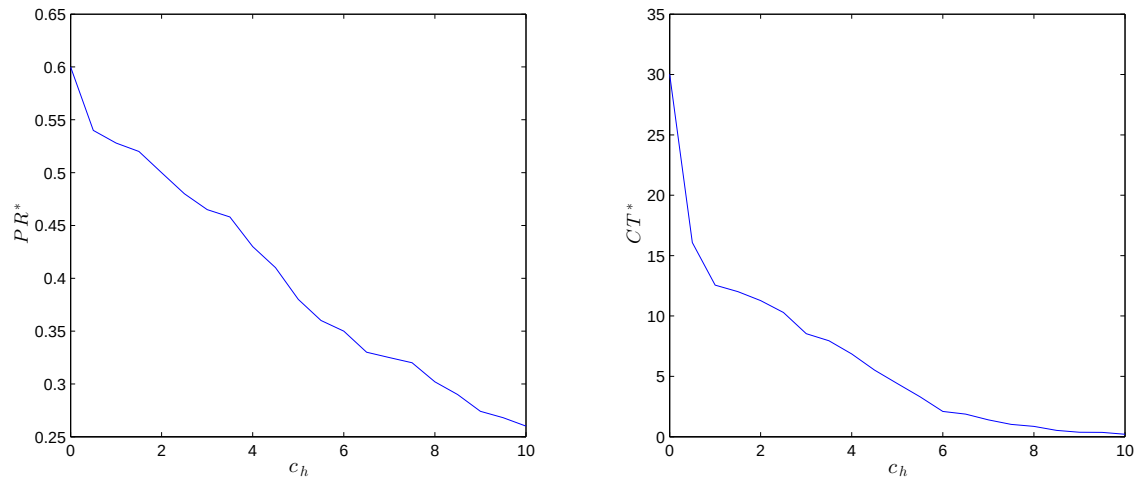


Figure 3.8: Optimal values of production rate and cycle time in the model with multiple up and down states and exponential transition times ($\mu_u = 2$, $\mu_d = 1$, $p = 0.1$, $r = 0.15$, $p' = 0.1$, $r' = 0.15$, $g = 0.1$, $h = 0.2$, $r_Q = 0.15$, $c_p = 50$)

obtained by the analytical results.

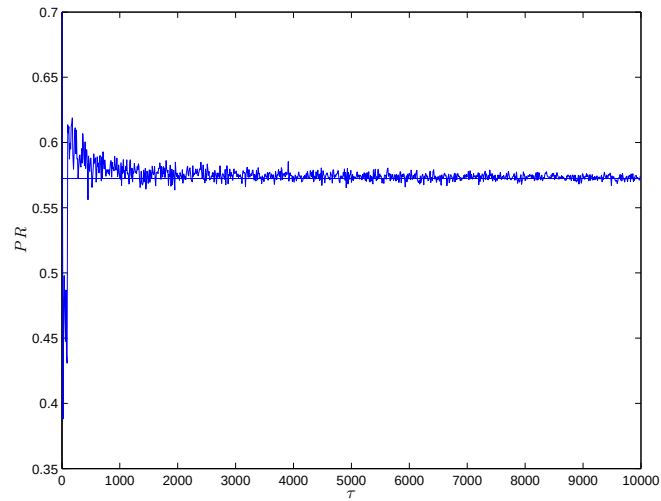


Figure 3.9: Convergence of the simulated production rate to the exact value in the model with multiple up and down states and exponential transition times ($\mu_u = 1$, $\mu_d = 1$, $p = 0.01$, $r = 0.1$, $p' = 0.05$, $r' = 0.1$, $g = 0.05$, $h = 0.1$, $r_Q = 0.1$, $Z = 10$)

Structure of the Mathematical Programming Problem:

Table 3.2 gives more detailed information about the structure of optimization based problem by defining the number of decision variables and the number of constraints for each run duration time. We investigated the dependency between the number of

Table 3.2: Structure of the optimization problem based on run duration time

Run Duration time (sec.)	Number of Discrete Variables	Number of Continuous Variables	Number of Equality Constraints	Number of Inequality Constraints	Optimization-based Simulation time (sec.)
10000	598	2374	1790	2959	0.2372
20000	1121	4526	3406	5647	0.5193
30000	1827	7384	5558	9211	1.4391
40000	2332	9432	7101	11764	4.01
50000	2867	11588	8722	14455	8.0314
60000	3557	14398	10840	17957	14.594
70000	4062	16414	12353	20476	20.764
80000	4800	19408	14609	24208	30.121
90000	5504	22243	16740	27747	34.071
100000	5857	23666	17810	29523	38.365
110000	6648	26858	20211	33506	46.351
120000	7306	29515	22209	36821	53.287
130000	8032	32459	24428	40491	59.667
140000	8443	34101	25659	42544	62.391
150000	9176	37087	27912	46263	70.865

discrete variables N_{disc} , number of continuous variables N_{cont} , the number of equality constraints N_{eq} and the number of inequality constraints N_{ineq} and the solution time $t_{solution}$ empirically by using the data given in Table 3.2. The following polynomial

equation is estimated by using multiple regression:

$$t_{solution} = 0.76(N_{disc})^2 - 0.60(N_{cont})^2 + 0.27(N_{eq})^2 - 0.33(N_{ineq}) + 0.32(N_{ineq})^2. \quad (3.70)$$

As discussed in Section 3.4.4, the computation time does not depend on the structure of the system such as the number of states or the buffer capacity. Our numerical results validate that the buffer capacity does not have any significant effect on the solution time.

Comparison of Discrete and Continuous Simulation and Mathematical Programming Approach:

In order to evaluate the computational efficiency of the iterative solution, the mathematical programming based solution, and the discrete event simulation of the discrete material flow approximation of the continuous flow, we use the same system given in Section 3.3.7. We use exponential transition times in order to compare the accuracy of simulation with analytical results. The same sample realizations for state transition times are used for all the methods.

Table 3.3 shows the results of this comparison study. The first column of this table gives the run duration time for the system. In the second column of the table, we used discrete time, discrete state space model with step size of $\Delta t = 0.01$ to simulate the system. The step size is chosen in a way that gives similar percentage error with respect to the analytical results for the same run length. In the third column, we simulated the system iteratively, solving the equations, which shows the dynamics of the problem using mathematical programming formulations from Section 3. In the last column, the system is simulated by optimization based model from Section 4. The percentage errors are also given in the table.

As shown in Table 3.3, continuous simulation based on mathematical programming is approximately 800 times faster than discrete simulation. As it was emphasized in previous sections, this method improves the speed of simulation by focusing just on the event times when we have a change in processing rate. The approach speeds

up the simulation by determining the length of the times that flow rate stays the same up to the next change due to any variation in the states of machines or reaching specific levels. Therefore it yields a more efficient performance evaluation compared to a simulation that uses discrete flow of materials to approximate a continuous flow. The fifth column of Table 3.3 indicates the simulation time using optimization based approach. For instance, when the run duration is equal to 150000 seconds, the optimization based formulation has 37087 continuous variables, 9176 binary variables, 27912 equality constraints and 46263 inequality constraints. This problem can be solved in 71 seconds. Based on all the results in fifth column, currently it takes a longer time to do optimization-based simulation compared to discrete simulation. However, it is an alternative representation for the system dynamics and this approach allows simultaneous optimization and simulation.

Table 3.3: Comparison of simulation times based on run duration time

Run Duration time (sec.)	Discrete Simulation time (sec.)	Percentage Error (%)	Continuous Simulation time (sec.)	Optimization-based Simulation time (sec.)	Percentage Error (%)
10000	2.95	8.54	0.0037	0.2372	8.42
20000	5.93	7.69	0.0079	0.5193	7.64
30000	8.17	6.62	0.0165	1.4391	6.67
40000	10.59	6.63	0.0192	4.01	6.61
50000	13.28	6.60	0.0233	8.0314	6.61
60000	15.90	6.27	0.0278	14.594	5.2
70000	17.83	6.27	0.0297	20.764	6.41
80000	22.38	5.84	0.0341	30.121	5.82
90000	22.96	5.84	0.0371	34.071	5.91
100000	25.51	4.08	0.0414	38.365	3.06
110000	28.11	3.03	0.0461	46.351	3.16
120000	30.85	3.82	0.0475	53.287	2.23
130000	33.32	2.46	0.0527	59.667	1.22
140000	35.77	1.78	0.0761	62.391	1.73
150000	45.37	1.32	0.1315	70.865	0.98

Chapter 4

MODELING AND ANALYSIS OF A COOPERATIVE SERVICE NETWORK

Technology has improved both the types of health services available, and how they are delivered. Telemedicine is one of these substantial advances, which leads to more convenient access for patients and utilization improvement for providers.

Telemedicine is defined as the use of medical information exchanged over a distance using means of IT to diagnose and treat patients' diseases. It includes a wide variety of applications and services using two-way video, email, smart phones, wireless tools, and other forms of telecommunications technology which enable the patients to access doctors through an easy to use on-line platform (Roine et al, 2001; Norris and Norris, 2002; Craig and Petterson, 2005; Whitten et al, 2010). Recent technological advances have enabled the introduction of a broad range of telemedicine applications, such as teleradiology, teleconsultation, telesurgery and remote patient monitoring that are supported by computer networks and wireless communication (Binkhuysen and Ranschaert, 2011; Sorknæs et al, 2011; Pawar et al, 2012; Peirce et al, 2011).

The use of a telemedicine platform enables the doctors to serve the health needs of patients more efficiently and allows the patients to avoid the cost and inconvenience of travel and time away from work and family responsibilities (Bynum et al, 2003; Gustke et al, 2000; Hailey et al, 2004; Wootton et al, 2011; Lemon et al, 2018). This is necessary given our aging population and the increase in the occurrence of chronic diseases. By utilizing telemedicine structure, physicians can serve more patients without the need to hire more staff or increase office space. The spread of disease is also reduced as individuals with contagious diseases do not have to expose it to others in crowded waiting rooms.

Currently, the use of telemedicine is becoming more widespread as the technology barriers have been overcome. NowClinic, Doctor on Demand, and American Well companies are a few examples which focus on a fairly straightforward brand of telemedicine: Letting patients confer with doctors over video.

It is very important to consider the effective design and management of telemedicine platform to offer the patients a desired service level, increase the utilization of the resources and maximize the profit. To this end, we present the analysis of the service network structure by using single queue approximation method. As a result, we obtain the explicit expressions for the system performance measures depending on the main parameters of the network. There are several papers on telemedicine platform performance evaluation from different perspectives (Scott and Mars, 2015; Chang, 2015). The main concern indicated in all these works is the patients' satisfaction which is mainly the result of the lower waiting time and higher service level. The main advantage of our offered model for the telemedicine platform is that there is no waiting time for the platform requests to receive the medical service because they are dispatched to one of the available doctors upon their request to start the service immediately. However, there is a probability of rejection if there is no available doctor in the network. Therefore, it is definitely essential to design the platform in a way to guarantee the desired service level for platform patients. Regarding different service level expectations, we discuss different designing parameters such as the number of participating doctors and their utilization due to their own patients.

In this part, our main objective is to answer the following main questions: *What is the optimal admission policy for each doctor in the platform to visit different groups of patients in a non-preemptive manner?*; and *How do the designers of telemedicine platform decide on different parameters of the telemedicine cooperative network?*

We present the structure of a telemedicine platform as a cooperative service network which contains a number of physicians to serve the patients asking for on-line medical service. The doctors receive the platform patients in addition to their own patients who come in person. The visiting service is a non-preemptive process, and the doctors are

able to admit a patient when they are idle. There is no waiting queue for the platform patients in this model, and they are dispatched to one of the available doctors in the platform immediately upon their on-line request. On the other hand, the doctors' own patients join a waiting queue upon their arrival, and they are served according to the FCFS rule. Hence, due to non-preemptive characteristic of the model, all of the doctors participating in the network should decide their optimal strategy to serve different groups of patients. In other words, they should decide whether to become available in order to receive the patients from telemedicine platform when their own patients are already waiting in the system.

The telemedicine platform that utilizes the cooperating doctors is expected to provide remote assistance for the patients from one site to another, and therefore, generates an additional stream of patients for the doctors to achieve their full potential in serving their patients. We analyze this model in two steps. In the first step we determine the optimal admission policy for each independent doctor in the platform for their own patients and platform patients. We model the doctors' decision problem as a Markov decision process with the objective of maximizing the total revenue of the doctors over a finite time horizon. We show that the optimal admission policy for the doctors in the network is a threshold type policy. Following this result, we evaluate the performance of the system by using a queueing model and show that the optimal policy is a strict non-preemptive priority for their own patients over an infinite horizon. In the second step, we evaluate the performance of the telemedicine network with a number of homogeneous doctors that use the optimal admission policy derived in the first step.

The main contribution of this part is two-fold. Firstly, we present a detailed operational model of the cooperative service network in the telemedicine platform setting and discuss the designing issues to obtain the benefits of the cooperation both for the patients and the doctors. Secondly, we prove that the optimal policy of a server that faces the demand from two different customer streams, one that join the queue and the other one that is lost, is a non-preemptive threshold type policy. None

of these issues have been addressed in the literature before.

4.1 Past Work

Much of the published works in the literature concentrates on the power of the technology and the content and quality of interactions it supports in telemedicine services (Hung and Zhang, 2003; Niyato et al, 2007; Devaraj and Ezra, 2011; Warren et al, 2011). However, there are a few number of works which focus on the operations management in telemedicine context (Rajan et al, 2018; Erdogan et al, 2018) and designing an efficient platform (Körpeoğlu et al, 2014). To fill this gap in the literature, we present an operational model of a telemedicine platform and discuss the designing points to obtain the efficient performance.

One of the important operational issues in health care management is the resource allocation between heterogeneous patient streams and admission decisions that should be made upon each arrival and service completions. We will give a brief review of past work related to dynamic admission control problem.

In different kinds of applications, a single server is usually shared by multiple types of customers with different revenue, arrival process and service time. So, there is a vast literature on defining the optimal dynamic admission control for revenue management in service and manufacturing systems.

Lippman and Ross (1971) is one the earliest works which analyze the optimal admission policy for a server with no waiting room, and the server receives the jobs according to a joint service time and reward probability distribution. Later on, Harrison (1975) shows that the $r\mu$ rule is optimal for scheduling a single server queue with two classes of jobs characterized different service rates μ_i , and rewards, r_i .

Örmeci et al (2001) consider the dynamic admission control problem in a Markovian loss queueing system with two classes of customers and show the optimality of threshold type admission policy in this structure. This model is also analyzed under different kinds of assumptions regarding the arrival process. Örmeci and Burnetas (2004) show the optimality of sequential threshold policy in the multi-class Markovian loss

system receiving random batches. Örmeci and van der Wal (2006) also establish the existence of optimal acceptance thresholds for all job classes in loss system with general interarrival times. Ku and Jordan (2003) consider multi-server loss queues in series and use a rate estimation technique for upstream stations, and a revenue estimation technique for downstream stations to obtain near optimal admission control. Righter (2000) and Rykov (2001) consider the admission control for the systems with heterogeneous servers. In terms of health care application, Pan et al (2017) use a finite-horizon dynamic programming approach to derive the optimal resource allocation strategy between early low priority patients and potential future higher priority patients in a hierarchical health system.

It is very important to emphasize that all these works consider the preemptive model and formulate the dynamic programming based on this assumption. There are also some works in the literature which use preemption as a control tool in their proposed model. Brouns and Van Der Wal (2006), Ulukus et al (2011) and Turhan et al (2012) are some examples for this kind of works.

However, the main difference of our work with the others existing in the literature is that the operation of each doctor is defined as a non-preemptive process. Moreover, we have different settings for different class of patients to join the system: own patients join to a queue and wait up to their admission time whereas the platform patients do not have any waiting queue, and they are lost if the doctor is busy at their request time.

There are a few number of researches dealing with non-preemptive models. Iravani et al (2007) consider the optimal single server scheduling in non-preemptive finite-population queueing system with heterogeneous customers. They show that the optimal service policy is a simple static priority policy for those classes that are served. Bispo (2011) studied single server scheduling problem with non-preemptive assumption, different classes of customers, and convex costs depending on arrival rates. He establishes a policy to reach near optimal performances, and presents it as a function of individual loads in the system. Zhao and Wang (2009) investigate the

performance of non-preemptive $M/M/1$ queueing system with two priority, and derive the accurate probability of the server which is busy or free in the system.

In this study, we change the standard formulation of resulting Markov decision process to accommodate our model where the doctor's state need to be captured in the overall state description. We derive results on the structure of the optimal admission policy for this problem and give an approximate analysis of the telemedicine network.

4.2 Model Description

In this study, we consider a telemedicine network which connects the patients to a number of homogeneous doctors through an on-line platform in order to provide health care service from a distance.

Telemedicine Platform. We assume that the entire number of doctors participating in the platform is equal to N . Following any request for a doctor examination through the platform, the on-line patient is assigned to one of the available doctors to start the service immediately. We call this group of patients *platform patients*, and assume that they place on-line requests according to a Poisson process with the rate of λ_t . The platform dispatches each of them to one of the available doctors in the network who is idle at the request time. Therefore, the platform patients do not wait in any queue to receive telemedicine service, and they are rejected, if there are no available doctors at their arrival time. The effective rate of platform patients' arrival for each doctor in the platform is defined by λ_e in this study. The platform patients pay a price of p_t for consultation with the doctor if they are accepted by the platform and each doctor makes a revenue of p_p by serving a platform patient ($p_p < p_t$). It means that the platform earns $p_t - p_p$ for each patient who receives the on-line telemedicine service. The long run proportion of the platform patients accepted, and served by the platform doctors is defined by γ_t . Thus γ_t is the service level offered by the platform for the on-line patients. The desired service level for the platform patients is also given by γ_d .

Doctors. The doctors in the platform serve their own patients as well. Each doctor

receives his own patients independently of the platform and according to a Poisson process with an arrival rate of λ_o . The patients of a doctor join a queue and wait to be visited by the doctor. These patients pay p_o , and they are served by the order of their arrival to the doctor. Each doctor incurs a holding cost of h for his own patients waiting in the queue. $\bar{L}_Q$ is the average number of the patients waiting in the queue, and it is dependent on the doctor's admission policy to serve the patients. Moreover, the visiting durations of all patients are independent and exponentially distributed with mean $1/\mu$ regardless of the patient group, and the number of the patients waiting in the system. The utilization of the doctors due to their own patients is defined by ρ . Once a doctor starts to serve a patient, the service will not be preempted, and upon conclusion of a service, the patient being served leaves the system. The offered service level by each doctor for the platform patients is the long run proportion of the platform patients served by the doctor and defined by γ_p .

Figure 4.1 shows the schematic of telemedicine platform along with the single doctor participating in the platform and receiving both groups of patients.

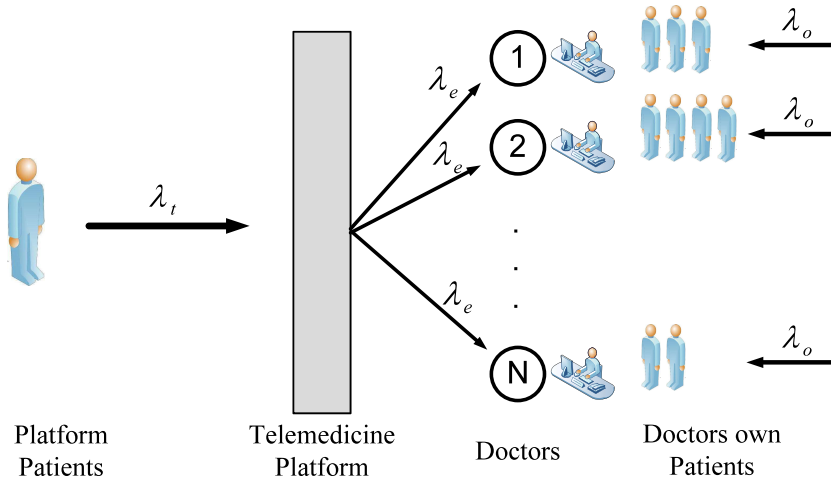


Figure 4.1: Telemedicine platform

4.2.1 Decision Problems

Effective management of this network is very critical because it is very important to offer the patients a high service level on one side and to increase the utilization and profitability of the doctors on the other side. We face with two main decision problems in this study. In the first step, we start with defining the optimal way of serving the patients for each member of the platform and, in the second step, we examine the platform design problem to reach the desired performance of the whole structure.

Doctors decision problem. The average total profit for each doctor in the platform (Π_D), is the profit which comes from his own patients (Π_o) plus the profit that comes with receiving platform patients (Π_p).

$$\Pi_D = \Pi_o + \Pi_p \quad (4.1)$$

and

$$\Pi_D = \lambda_o p_o + \gamma_p \lambda_e p_p - h \bar{L}_Q. \quad (4.2)$$

In this equation, the first term shows that each doctor makes a revenue of p_o from an arrival of his own patients. The second term shows that the doctor makes a revenue of p_p whenever he accepts a platform patient. The value of the offered service level for platform patients, γ_p , depends on the doctors' decision about the way of serving platform patients. Note that the value of exact state-dependent arrival rate of platform patients, λ_e , also depends on the doctor's admission policy. Finally, the last term depicts the waiting cost of the patients in the queue.

Therefore, we need to know the optimal way of serving different groups of patients for each member before going through the platform level designing problem.

Network decision problem. We examine the operation of the platform, and some of the designing points in the second step of our analysis. The average profit for the telemedicine platform, (Π_t), is defined as follows:

$$\Pi_t = \gamma_t \lambda_t (p_t - p_p). \quad (4.3)$$

It is important for the platform designers to offer a desired service level for the patients, ($\gamma_t \geq \gamma_d$), and maximize the average profit by setting the price p_p to the minimum possible value. All these parameters are dependent on the number of members in the platform and their utilization due to their own patients. Therefore, we consider the platform design problem where the decision variable is the number of doctors in the platform, and discuss the problem for different parameters of the system.

4.3 Analysis of the Doctors' Admission Policy for Platform Patients

The system evolves as a continuous-time Markov chain, and each doctor in the platform should decide on the policy that he will serve different types of patients depending on the state of the system and not on the prior history. Hence, we employ a Markov decision process framework to find the optimal control policy. The objective is to obtain dynamic scheduling policy of the doctors to maximize the discounted profit with discount parameter β in a finite time horizon.

We define $x(t)$ as the number of the doctor's patients present in the system at time t . The overall state description needs to capture the server state, because we consider non-preemptive model. Therefore, we define the state of the system as $S(t) = (x(t), j(t))$, where $j(t) \in \{0, 1\}$ is the state of a doctor at time t . If $j(t) = 0$, the doctor is idle, and $j(t) = 1$ means that the doctor is serving a patient. In this model, the only decision points for the doctors are arrivals of own patients when they are idle and the conclusion of a patient service.

We use the standard tools of uniformization to convert the continuous time Markov process into a discrete-time equivalent. omitting the explicit time dependency, we assume that the time between two consecutive transitions is exponentially distributed with rate $\delta = \mu + \lambda_o + \lambda_e + \beta$, where λ_e is the effective arrival rate of platform patients for each doctor. Without loss of generality, $\delta = 1$. Therefore, the value iteration algorithm for the dynamic scheduling of each doctor becomes:

$$v_{n+1}(x, 0) = \lambda_o \max \left\{ v_n(x+1, 0) + p_o, v_n(x+1, 1) + p_o \right\} + \lambda_e \left[v_n(x+1, 1) + p_p \right] + \mu v_n(x, 0) - hx \quad \forall x$$

$$(4.4)$$

$$v_{n+1}(x, 1) = \lambda_o \left[v_n(x+1, 1) + p_o \right] + \lambda_e v_n(x, 1) + \mu \max \left\{ v_n(x-1, 0), v_n(x-1, 1) \right\} - h(x-1) \quad \forall x \geq 1 \quad (4.5)$$

The doctors are available to receive platform patients when they are idle. According to Equation (4.4), when the doctor is idle ($j(t) = 0$), the maximization operation corresponds to the own patient's arrival to the queue. In this point, the doctor should decide whether to start serving the own patient from the queue or stay idle to receive next platform patient's on-line request. When a platform patient arrives in this case, the doctor starts his service immediately (the second term). In the third term, we have the service completion rate which does not have any effect on the state of system when the doctor is idle. Finally the last term in Equation (4.4) is the holding cost of the own patients waiting in the queue to receive service.

Due to the non-preemptive feature of the model, when the doctor is serving a patient ($j(t) = 1$), he is not available to accept the platform patients. According to Equation (4.5), the only decision point is the conclusion of the patient's service in this case. In this point, the doctor should again decide whether to start serving his own patient from the queue or stay idle to wait a platform patient's request.

4.3.1 Structure of the Optimal Scheduling Policy

Intuitively, we expect that it should be less profitable for the doctors to stay idle and accept platform patients when there are many own patients already in the queue waiting for the service. In this section, we will show the existence of an optimal threshold policy to maximize the expected β discounted reward over finite horizon using induction.

There are some results in the literature which guarantee the existence of an optimal threshold policy for specific systems where the model does not capture the state of the server but the number of jobs existing in the system (Altman et al, 1998; Koole, 1998; Örmeci et al, 2001). They show that the benefit due to an additional job in a

specified class is decreasing in the number of jobs from other classes. We will modify these results for our model to show the optimality of threshold policy as follows:

$$v_n(x, 1) - v_n(x + 1, 1) \leq v_n(x, 0) - v_n(x + 1, 0) \quad \forall x \geq 1; \quad (4.6)$$

$$v_n(x + 2, 1) - v_n(x + 1, 1) \leq v_n(x + 1, 1) - v_n(x, 1) \quad \forall x \geq 1; \quad (4.7)$$

$$v_n(x + 2, 0) - v_n(x + 1, 0) \leq v_n(x + 1, 0) - v_n(x, 0) \quad \forall x; \quad (4.8)$$

$$v_n(x + 1, 1) \leq v_n(x, 1) \quad \forall x \geq 1; \quad (4.9)$$

$$v_n(x + 1, 0) \leq v_n(x, 0) \quad \forall x. \quad (4.10)$$

Inequalities (4.9) and (4.10) show that when a new patient enters the system, the expected discounted profit decreases. It means that an additional patient incurs a positive opportunity cost. Moreover, Inequality (4.6) implies that the opportunity cost of having an additional patient when the doctor is idle is always greater than having an additional patient when the doctor is busy. The intuition behind this is the the optimality of threshold policy for each doctor in the platform. Inequalities (4.7) and (4.8) imply the concavity property of $v_n(x, 1)$ and $v_n(x, 0)$. We need to show that $v_n(x, 0)$ and $v_n(x, 1)$ are concave and decreasing in x to prove the optimality of the threshold policy in Inequality (4.6).

Theorem 4.1 *The optimal scheduling policy for the doctors participating in telemedicine platform is a threshold policy. Hence, there exists a threshold level R , such that it is optimal to stay idle, and accept arriving platform patients if the total number of patients already in the system is under threshold level R . Consequently, it is optimal to serve only the own patients, and reject the platform requests otherwise.*

The proof for this theorem is given in Appendix C provided in Chapter 6.

4.3.2 Stationary Analysis of the System

In the previous section, we show that under a discounted profit criterion, the optimal control policy for the doctors can be specified by a threshold level. In this section,

we seek to compute the average profit of the doctors under such a policy. Explicit expressions for the average profit can be obtained by analyzing a related queueing model. To obtain the queueing model for each doctor in the network, we need to modify the state space which shows the number of doctor's own patients in the system and also the existence of platform patients. The state of the system is defined by (x_1, x_2) , where x_1 indicates the number of doctor's own patients in the system and $x_2 \in \{0, 1\}$ shows whether there is a platform patient in service ($x_2 = 1$) or not ($x_2 = 0$).

Figure 4.2 shows the state space diagram of the resulting queueing model for the doctor's operation. According to this diagram, each doctor in the platform serves only the received platform requests if the number of the waiting patients is less than the threshold level, $(x_1 \leq R - 1)$. In consequence, the doctor starts to serve his own waiting patients whenever the number of them reaches the threshold level.

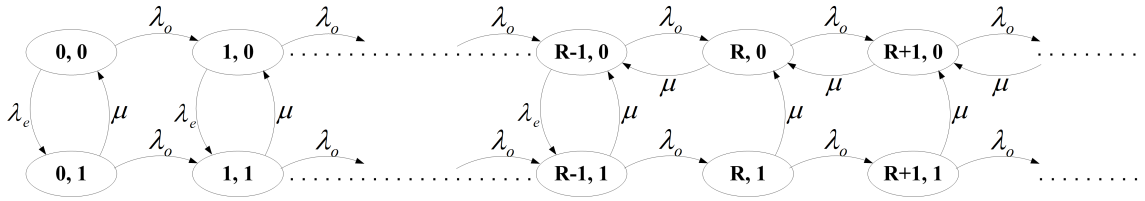


Figure 4.2: The state space diagram of queueing model for doctor's operation

Using the steady-state queue length probabilities, we can write the profit rate function in Equation (4.2) as follows:

$$\Pi_D = \lambda_o p_o + \lambda_e Pr(x_1 < R, x_2 = 0) p_p - h \bar{L}_Q. \quad (4.11)$$

The second term shows that the doctor makes a revenue of p_p whenever the number of waiting own patients in the queue is less than the threshold level R and there is not any platform patient already in the system. $Pr(x_1 < R, x_2 = 0)$ defines the steady state probability in this case and we can interpret the value of $Pr(x_1 < R, x_2 = 0)$ as the service level offered by each doctor for the platform patients, γ_p .

Calculating the steady state probabilities of the structure yields:

$$Pr(x_1 < R, x_2 = 0) = \frac{1 - \rho}{1 - \lambda_e/\mu} \quad (4.12)$$

Equation (4.12) indicated that the steady state probability of accepting platform patients does not depend on the value of threshold level R . The reason for this result is that all the states under the threshold value are transient states and the value of γ_p is equal to $Pr(x_1 = R - 1, x_2 = 0)$ and independent of all the states smaller than $R - 1$ ($x_1 < R - 1$).

Due to the independence feature of Equation (4.12) from the value of R , we should minimize the average number of own patients waiting in the queue ($\bar{L}_Q$) to maximize the average long run profit obtained by each doctor in the platform. It is quite intuitive that the value of $\bar{L}_Q$ is increasing in the value of R . Therefore, setting $R = 1$ will maximize the profit and leads to the following formulation for $\bar{L}_Q$:

$$\bar{L}_Q = \frac{\lambda_o(\rho + \lambda_e/\mu)}{\mu - \lambda_o + \lambda_e(1 - \rho)} \quad (4.13)$$

and the value of γ_p becomes equal to $Pr(x_1 = 0, x_2 = 0)$. This observation yields the following proposition on the non-preemptive priority policy.

Proposition 4.1 *The stationary optimal scheduling policy for the doctors in the platform is a non-preemptive-priority policy which implies that the doctors serve their own patients whenever there is at least one patient in the queue in a non-idling manner. It means that the threshold level is equal to one in this case ($R = 1$) and the doctors are available to receive the platform requests whenever there is not any patient in the queue.*

The proof for this proposition is given in Appendix C provided in Chapter 6.

4.4 Analysis of the Telemedicine Platform under the Optimal Policy

In this section, we will analyze the telemedicine platform operation. According to the optimal scheduling policy for each of the doctors as the members of the platform in

Section 4.3.2, we now focus on designing the parameters of the whole structure for an efficient operation.

The focus of the platform is on ensuring a desired service level for the patients; this is usually a major concern in service systems. Therefore, the main designing parameters could be defined as the offered price for the platform patients, utilization of the participating doctors (due to their own patients) and the number of the members in the platform.

The cooperation in the telemedicine platform should provide financial benefits for the doctors to convince them to join the network. The doctors should make sure that they will obtain more profit by entering to the platform and serving on-line patients compared to the case that they serve only their own patients independently. Regarding the independent operation of the doctors with just their own patients, we face an $M/M/1$ queueing model where the patients join the queue with the rate of λ_o , pay the price of p_o , and get the service with the rate of μ . Hence, we can derive the minimum value of the price p_p by ensuring that the long run average profit for each member of the platform in Equation (4.11) is always greater than or at least equal to the resulting long run average profit of the doctor's independent operation. As the result, the doctors accept to work in the telemedicine platform if and only if:

$$p_p > h \frac{\rho}{\mu - \lambda_o}. \quad (4.14)$$

It means that each platform patient must bring at least the revenue of $h \frac{\rho}{\mu - \lambda_o}$ for each member to work in the telemedicine platform and accept platform's on-line patients.

From the platform designers perspective, It is not beneficial to define p_p larger than its minimum possible value in Equation (4.14), because it does not have any effect on the service level offered by the doctors for the platform patients (γ_p), and the larger values of p_p lead to the lower profit level for the telemedicine platform as defined in Equation (4.3).

In this study, the platform receives the patient on-line requests and dispatch them among available doctors with the same probability due to homogeneity of the participating doctors. Hence, the resulting queueing network can be analyzed by using

the Single Queue Approximation (SQA) method (Hosseini and Tan, 2019).

The main idea behind SQA is analyzing each queue in the queue network independently from the others and approximating the steady-state queue length probabilities by capturing the effect of other queues through modeling the arriving process by a stochastic process with state dependent rates. It means that we can obtain all the performance measures of the platform by analyzing just a single doctor if we can model the effective arrival rate of the platform patients, λ_e , depending on the state of other doctors in the platform.

Let us assume that when a platform patient is dispatched to one of the doctors, there are also k idle doctors among the other $N - 1$ members of the platform. Since it is equally likely for all the available doctors to receive the arriving patient, the approximate long-run rate for receiving platform patient λ_e is

$$\lambda_e \simeq \lambda_t \left(\sum_{k=0}^{N-1} \frac{1}{k+1} Pr(K = k) \right) \quad (4.15)$$

where K is a random variable that shows the number of idle doctors who are available to receive platform requests. Since the steady-state probability of satisfying platform patients is γ_p , $Pr(K = k)$ can be approximated by using the binomial distribution:

$$Pr(J = j) = \binom{N-1}{k} (\gamma_p)^k (1 - \gamma_p)^{N-k-1}. \quad (4.16)$$

Equations (4.15) and Equation (4.16) yield the following closed-form approximation for λ_e :

$$\lambda_e \simeq \lambda_t \frac{1 - (1 - \gamma_p)^N}{N \gamma_p}. \quad (4.17)$$

We can also write the approximate value of the platform service level, γ_t , as follows:

$$\gamma_t \cong 1 - (1 - \gamma_p)^N. \quad (4.18)$$

Therefore, if the aim of designing the platform structure is to offer the patients a desired service level, γ_d , then we should consider the utilization of the working doctors, the number of participating members and the effective arrival rate for each member

as the result of servers' utilization and total arrival rate of platform patients. All of these parameters are highly correlated with each other and we are going to discuss different aspects of these relations in the numerical experiments in the next section.

4.5 Numerical Experiments

Four key principles of the platform design are the number of participating doctors (N), their utilization (ρ), the platform desired service level (γ_d) and its patients' arrival rate (λ_t). In this section, we are going to discuss these parameters of the platform which are all important in designing the structure to assure a desired user experience while leading to the platform and its members profitability.

If the designers have knowledge about the doctors' utilization, they can decide on the number of members to guarantee the desired platform service level. For example, the doctors in a selected segment can have similar utilization and the platform can invite the doctors accordingly. Figure 4.3 refers to this general case, and shows the number of participants for different values of doctors' utilization and desired service level. According to this figure, the higher utilization values of the members in the

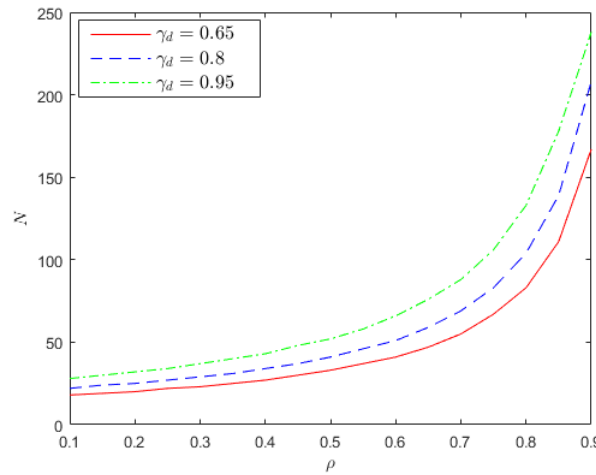


Figure 4.3: The number of participating doctors in the telemedicine platform (N) for different values of doctors' utilization (ρ) and platform desired service level (γ_d). ($\mu = 1, \lambda_t = 25$)

telemedicine platform leads to the higher number of participants in order to satisfy the desired service level for platform users.

It is also possible to look at the problem from the opposite perspective. Figure 4.4 represents the utilization value for the doctors based on the number of participants and the desired service level. This graph could be used when the designers face with some

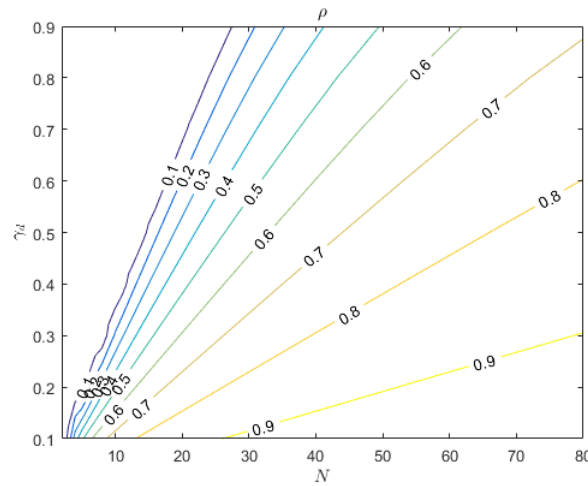


Figure 4.4: The utilization values of participating doctors in the telemedicine platform (ρ) for different numbers of members (N) and platform desired service level (γ_d). ($\mu = 1, \lambda_t = 25$)

limits regarding the number of participants and try to choose the doctors with suitable value of utilization to ensure the desired service level. In this case, the effective arrival rate of the platform patients for each doctor and their offered service level are also given in Figure 4.5.

We can also show the effective arrival rate of platform patients (λ_e) for each doctor in the network based on the doctors' utilization and their offered service level for the platform patients (γ_p) in Figure 4.6. According to Figure 4.6, whenever the doctors' offered service level for the platform patients is low, they receive platform requests with a higher rate during their available period and the reason is the lower number of available doctors in the platform.

The most important aspect in designing the telemedicine platform is to serve the

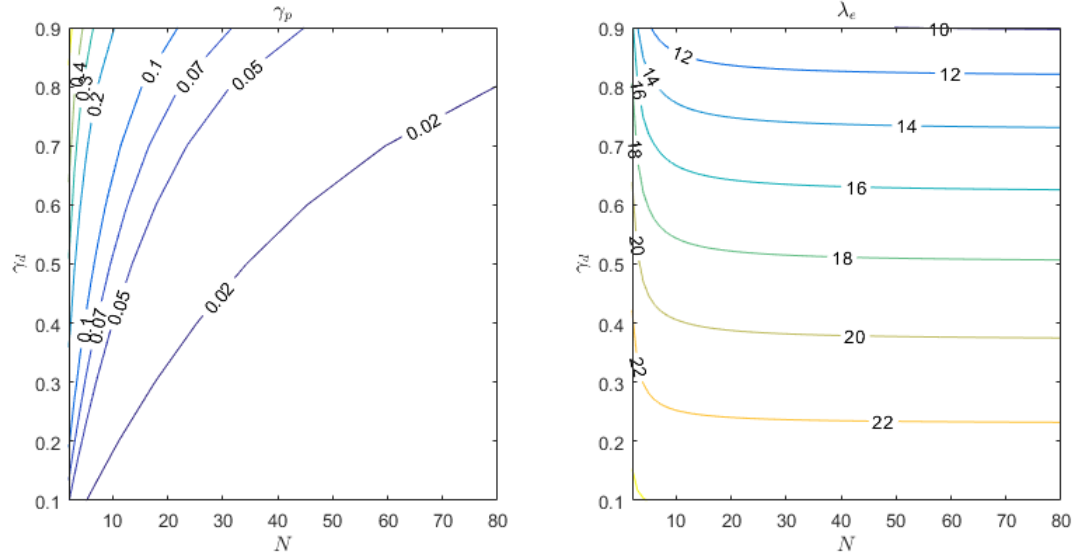


Figure 4.5: The offered service level for the platform patients by each doctor in the telemedicine platform, γ_p and the effective arrival rate of platform patients, $\lambda_e(\mu = 1, \lambda_t = 25)$

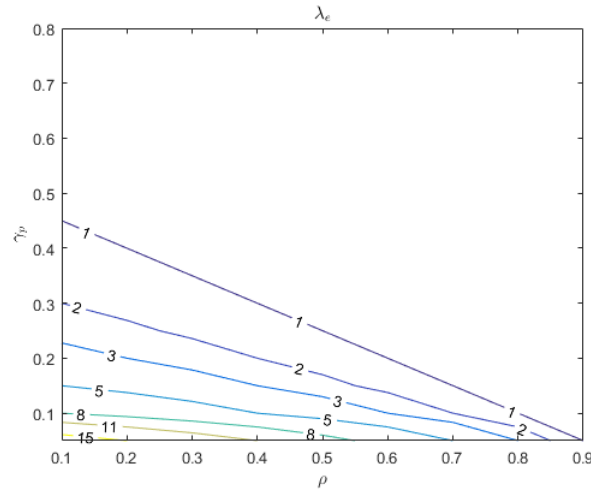


Figure 4.6: The effective arrival rate of platform patients for each doctor in the network based on the doctors' utilization and their offered service level for the platform patients. ($\mu = 1$)

patients with a high service level and the main uncertainty is the patients' on-line request rate which may be subject to change due to many different reasons. Hence,

it is essential to show that how the number of members in the platform should be changed in response to the fluctuation of λ_t to give the same service level for platform patients. Figure 4.7 shows this effect for three different values of members' utilization. According to figure 4.7, whenever the members of the platform have higher utilization,

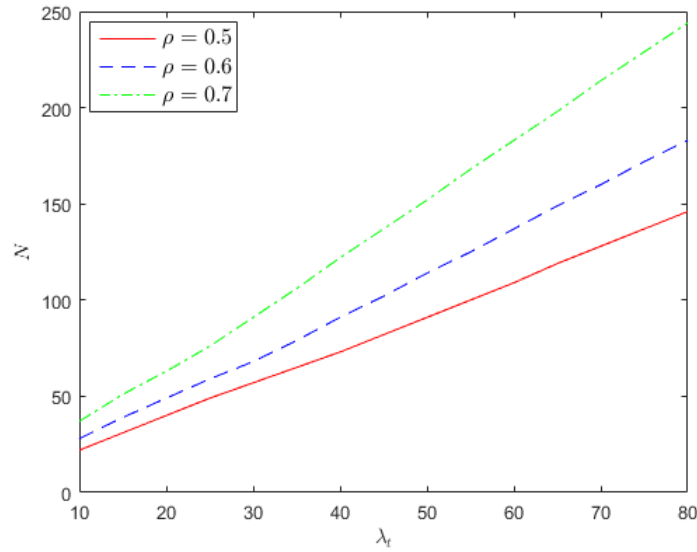


Figure 4.7: The number of participating number of doctors in the platform for different values of platform patients request rate and for three different values of doctors' utilization ($\mu = 1, \gamma_d = 0.9$)

we should increase the number of participants with a higher rate to satisfy the desired service level for the platform patients.

We can also consider the changes in the platform offered service level (γ_t) as the result of fluctuations in the value of λ_t in the telemedicine platforms with different number of participants. Figure 4.8 shows that the small size telemedicine networks with lower number of members are more sensitive to changes in on-line requests rate than the large ones with a higher number of members.

In summary, We observe that the networks with high utilization level of participating doctors or lower number of members, show higher sensitivity to fluctuating platform patients request rate, and their offered service level decreases with a higher rate in response to increasing on-line request rate.

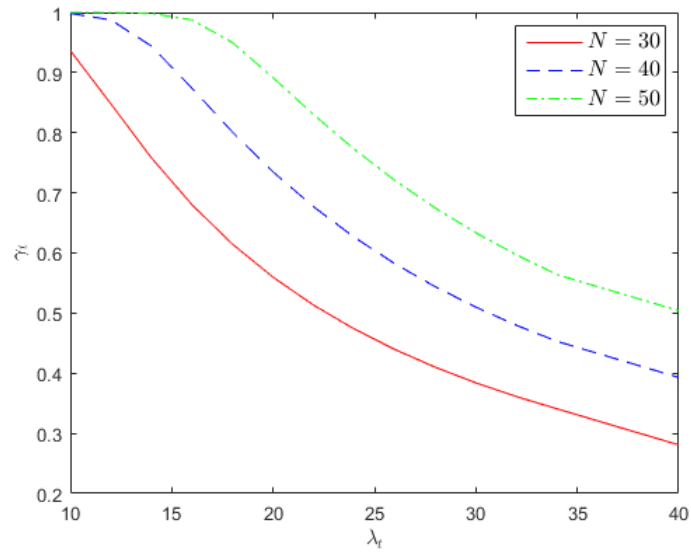


Figure 4.8: The platform offered service level for different values of platform patients request rate and for three different numbers of participating members, N ($\mu = 1, \rho = 0.6$)

Chapter 5

CONCLUSION

This dissertation is consisted of three parts. In the first part we investigate the operation of a cooperative manufacturing network, and discuss the benefits of the cooperation based on different parameters of the system. In the second part, we use mathematical programming approach to perform joint simulation and optimization of general continuous flow production systems. In the last part, we study the model of cooperative service network and consider telemedicine platform as an example for this kind of structure.

In the first part, we present an analytical model of a cooperative network among a number of manufacturers that want to increase their utilization and profit as a result of this cooperation. Our analysis is performed on two types of cooperative network with static and dynamic routing policies. We modeled the static network using $M/M/1$ queueing model and defined the optimal shares of each member from the network external customers. In a dynamic network we determine the optimal production, rationing and routing policies using mathematical programming to solve the resulting Markov decision problem. Our results show that a cooperative network with a dynamic routing policy yields a significant financial improvement compared to the static network. Additionally, using a dynamic routing policy leads to a higher improvement for a network with heterogeneous members compared to a network with homogeneous members, since the dynamic routing policy can react to changes in the inventory levels of heterogeneous producers effectively. We also show that for a dynamic network with identical members, the optimal routing policy follows the JSQ rule and based on this observation, the level of inventories are changing in a very similar way to each other. As the result, we suggest an approximation analysis of this

kind of network with a lower computational effort. We prove that we can approximate this structure just by analyzing a single queue in the network by using the single queue approximation. Using this procedure, we found the steady state probabilities of queue length to analyze the system performance. Our results show the high accuracy of our approximation method compared to analytical results.

Our results that are based on the analytical model show that a cooperative production network increases the producers profit especially when the external customers require a high service level. In addition, the main benefit of a cooperative production network is in its ability to provide a significantly higher service level to customers compared to the service level these customers would get from the producers independently. In the case of mass collaboration, such as the case in cloud manufacturing, the network can provide close to 100% service level and accept external customers even when they offer a very low price. As a result, the possibility of offering a high service level and low price will give a competitive advantage to the members of the cooperative network.

Moreover, we propose a mathematical programming formulation to define the best coalition structure among heterogeneous manufacturers. The optimal coalition scheme between the participating producers may involve multiple interoperable networks to maximize the total profit while ensuring that each producer is better off compared to the independent operation. Our results show that the optimal coalition structure may be different for the networks with static and dynamic routing policies.

The model for the cooperative production network presented in the first part can be extended to derive optimal rationing policies for different customer classes with different service level expectations. As another direction of this research, we can consider the competition among participating producers by using game theoretical approaches. These are left for future research.

In the second part, we present an analytical simulation method to determine the performance of general continuous flow systems with two stations with different modes of operation and different flow rates associated with each mode separated with a finite

buffer and general transition time distributions. Our method uses sample realizations of the station states as its input and generates a sample realization of the buffer in an efficient way by evaluating the system at each critical time where there is a change in flow rate due to several reasons such as any change in stations' status, flow rates, or reaching full or empty buffer levels.

As a result, our representational model reduces computation load dramatically in comparison to discrete event representation. The mathematical programming approach embeds the equations that describe the dynamics of the system to determine the solution of these equations as a solution of a mixed integer linear program. We show that the optimization of the mathematical program gives the unique solution of the set of equations that describe the dynamics of the system. This alternative representation allows both determining a simulated sample realization of the system as well as joint optimization and simulation of the system when the system parameters are to be determined to optimize a given objective function subject to additional constraints. To present this, we analyze the buffer capacity determination problem with the objective of determining the minimum buffer capacity that achieves a desired production rate and also with the objective of maximizing the profit.

The mathematical programming representation can also be used to analyze certain properties of the performance measures and their dependence on the system parameters. In this paper, we show the monotonicity property of the production rate with respect to the buffer size by using the mathematical programming representation.

Moreover, we investigated the computational performance of the proposed method in detail, and show that the number of decision variables and constraints depend on the rates of change among different states of the system. As a result, the computational effort of this approach does not depend on the number of states for each stage or on the buffer capacity.

Compared to using discrete event simulation that approximates continuous flow with discrete flow of materials, the analytical simulation method we present gives approximately 800 times faster results by just evaluating the system dynamics at the

critical times. The optimization-based approach also gives acceptable computational performance to evaluate the performance of a given system. Hence, combined with its flexibility to handle a wide range of systems and ability to optimize and simulate simultaneously, the proposed approach is a promising method for performance evaluation of continuous flow systems with general transition time distributions.

The work on using the mathematical programming to simulate and optimize production systems presented in the second part can be extended in a number of ways. The two station system that is considered can be extended to systems with multiple thresholds that are used to model batching machines and also that are used as a building block for decomposition of systems with loops. The model with multiple thresholds can also be used to analyze the flow rate control problem with multiple demand sources corresponding to different customer classes. When the thresholds are decision variables, the optimal values of these thresholds can be determined with this approach. We can also consider the multi-station systems in series or in general topologies with multiple buffers. With this extension, mathematical programming representation can be used to solve the buffer allocation problem. Finally, analyzing systems where the state-transition rates are affected by the buffer process, such as models with operational failures where failure rate is proportional to the processing rate requires an extension of the general approach to capture system dynamics. These are left for future research.

In the last part, we study the operational model of a telemedicine platform, as a collaborative service network among a number of doctors, to serve the on-line platform patients in addition to the doctors' own patients. To this end, firstly, we formulate a MDP framework to solve the dynamic admission problem for the members of the platform who face with different types of patients, and should manage their availability to receive platform requests. We modify the standard formulation of MDP to capture the doctors' state due to non-preemptive characteristic of the system, and explore the structure of the optimal scheduling policy by maximizing the value function over a finite time horizon. Our results show that the optimal scheduling policy for the

doctors participating in telemedicine platform is a threshold policy. Thus, there exists a threshold level such that it is optimal to stay idle, and accept only arriving platform patients if the whole number of patients already in the system is under threshold value.

By conducting the stationary analysis of the system using queueing model, we obtain the optimal control policy to maximize the average total profit over an infinite time horizon as well. We conclude that the stationary optimal scheduling policy for the doctors is to become available for the platform requests whenever there is not any patient already waiting in the queue.

Assuming that the doctors are using the optimal control policy to manage patients' admission, we perform the approximate analysis of the platform structure in the second step. Similar to the first part related to manufacturing network, we use single queue approximation method to obtain steady state probabilities, and analyze the performance of the telemedicine platform.

Finally, we discuss the platform design problem to define the optimal number of doctors including in the platform through numerical experiments. We perform our experiments for different values of members' utilization and platform desired service level, which are the most important parameters to evaluate the performance of the platform. We also show the effect of increasing telemedicine requests on the optimal number of members to satisfy service level expectation.

Due to our results, the networks with high utilization level of participating doctors, show higher sensitivity to fluctuating platform patients request rate, and their offered service level decreases with a higher rate in response to increasing on-line request rate.

The model for the cooperative service network investigated in the third part can be extended to examine the platform with heterogeneous types of doctors with different utilization levels and specialties. Similarly, different types of patients with different service rates and priorities can be analyzed in this structure. The model can also be extended to consider a waiting queue for telemedicine requests, and to derive optimal dispatching policy for the platform to improve the performance. These are left for

future research.

Overall, the detailed stochastic models for cooperative production and service networks allow us to understand the operation of these networks better and design them accordingly to achieve the maximum benefit. Modeling and analysis of these networks require determining the optimal admission and dispatching policies to assign different demand streams to different producers or service providers. In order to determine and analyze these policies, mathematical programming is used together with the stochastic models.

We believe that the results derived in this dissertation will be of value for the producers and service providers planning to join the cooperative networks. Our results help them and the network designers to understand the dynamics of these structures and identify the conditions and factors to reach the full benefits of cooperation. This dissertation also extends the state of art in the analyses of cooperative networks by considering different kinds of settings that have not been addressed in the literature. Therefore, the analysis regarding cooperative networks provides researchers a basis for further research and development on the new cooperation based business models both for the production and service systems.

Chapter 6

APPENDICES

6.1 Appendix A

6.1.1 Performance Measures of the Independent Operation when the External Business Offer Is Accepted

Case 1: ($\rho_i = 1$ and $\xi_{nc_i} \neq 1$)

$$Pr(I_{nc_i} > R_{nc_i}) = 1 - \frac{(R_{nc_i} + 1)\lambda_{ei}\xi_{nc_i}^{S_{nc_i}-R_{nc_i}}}{(\mu_i + \lambda_{ei} + \lambda_{ei}R_{nc_i})\xi_{nc_i}^{S_{nc_i}-R_{nc_i}} - \mu_i}, \quad (6.1)$$

$$Pr(I_{nc_i} > 0) = 1 - \frac{\lambda_{ei}\xi_{nc_i}^{S_{nc_i}-R_{nc_i}}}{(\mu_i + \lambda_{ei} + \lambda_{ei}R_{nc_i})\xi_{nc_i}^{S_{nc_i}-R_{nc_i}} - \mu_i}, \quad (6.2)$$

$$\begin{aligned} \bar{I}_{nc_i} = E[I_{nc_i}(t)] = & \quad (6.3) \\ & \left(\frac{\xi_{nc_i}^{S_{nc_i}-R_{nc_i}+1} - (S_{nc_i} - R_{nc_i} + 1)\xi_{nc_i} + S_{nc_i} - R_{nc_i}}{(1 - \xi_{nc_i})^2 \xi_{nc_i}^{S_{nc_i}-R_{nc_i}}} + \frac{(\xi_{nc_i}^{S_{nc_i}-R_{nc_i}+1} - 1)R_{nc_i}}{\xi_{nc_i}^{S_{nc_i}-R_{nc_i}}(\xi_{nc_i} - 1)} + \frac{R_{nc_i}^2 - R_{nc_i}}{2} \right) \\ & \times \frac{\lambda_{ei}\xi_{nc_i}^{S_{nc_i}-R_{nc_i}}}{(\mu_i + \lambda_{ei} + \lambda_{ei}R_{nc_i})\xi_{nc_i}^{S_{nc_i}-R_{nc_i}} - \mu_i}. \end{aligned}$$

Case 2: ($\rho_i \neq 1$ and $\xi_{nc_i} = 1$)

$$Pr(I_{nc_i} > R_{nc_i}) = \frac{(\rho_i - 1)(S_{nc_i} - R_{nc_i})}{\rho_i^{R_{nc_i}+1} + (\rho_i - 1)(S_{nc_i} - R_{nc_i}) - 1}, \quad (6.4)$$

$$Pr(I_{nc_i} > 0) = 1 - \frac{(\rho_i - 1)\rho_i^{R_{nc_i}}}{\rho_i^{R_{nc_i}+1} + (\rho_i - 1)(S_{nc_i} - R_{nc_i}) - 1}, \quad (6.5)$$

$$\begin{aligned} \bar{I}_{nc_i} = E[I_{nc_i}(t)] = & \quad (6.6) \\ & \left(\frac{\rho_i^{R_{nc_i}+1} - (R_{nc_i} + 1)\rho_i + R_{nc_i}}{\rho_i^{R_{nc_i}}(1 - \rho_i)^2} + \frac{S_{nc_i}(S_{nc_i} + 1) - R_{nc_i}(R_{nc_i} + 1)}{2} \right) \\ & \times \frac{(\rho_i - 1)\rho_i^{R_{nc_i}}}{\rho_i^{R_{nc_i}+1} + (\rho_i - 1)(S_{nc_i} - R_{nc_i}) - 1}. \end{aligned}$$

□

6.1.2 Performance Measures of the Independent Operation when the External Business Offer Is Rejected

In the model with two classes of customers, when the rationing level is equal to the base stock level, all the external customers are rejected. Hence, when the external business offer is rejected, we set $S_{nc_i} = R_{nc_i}$ and $\rho_i = \xi_{nc_i} = \lambda_i/\mu_i$ in Equations (2.2) - (2.4). These substitutions yield

$$Pr(I_{nc_i} > 0) = \begin{cases} 1 - \frac{\rho_i^{S_{nc_i}+1}(\rho_i-1)}{\rho_i^{S_{nc_i}+1}-1} & \rho_i \neq 1 \\ 1 - \frac{1}{S_{nc_i}+1} & \rho_i = 1 \end{cases}, \quad (6.7)$$

$$\bar{I}_{nc_i} = E[I_{nc_i}(t)] = \begin{cases} \frac{\rho_i^{S_{nc_i}+1} - (1+S_{nc_i})\rho_i + S_{nc_i}}{(\rho_i-1)(\rho_i^{S_{nc_i}+1}-1)} & \rho_i \neq 1 \\ \frac{S_{nc_i}}{2} & \rho_i = 1 \end{cases}. \quad (6.8)$$

6.1.3 Proof of Proposition 2.1

The proof is similar to the one which is used in Gupta et al (2007). We consider the case where there are two producers ($N = 2$). The result for $N > 2$ is a straightforward expansion of this result and it is not given here.

Based on our assumptions about the system, the system is modeled as a continuous time Markov chain where the state of the system is a two-dimensional vector representing the number of items at each inventory. Let $\pi_{n+1,j}$ be the steady state probability when there are $n+1$ items in the first inventory and j items in the second one. The steady state probability of having n items in the first inventory, P_n , can be written as follows:

$$P_{n+1} = \sum_{j=0}^S \pi_{n+1,j}. \quad (6.9)$$

The customer arrival rate to the first producer that is the arrival rate of customers depending on the number of items in the inventory (λ_{n+1}) can be defined as:

$$\lambda_{(n+1)} = \sum_{j=0}^S (\pi_{n+1,j}/P_{n+1})(\lambda_1 + \lambda_e q_{(n+1,j)}) \quad (6.10)$$

where $q_{n+1,j}$ is the probability that the router routes a demand to the first inventory when there are $n + 1$ items in the first inventory and j items in the second inventory. For two participants, we have:

$$q_{(n+1,j)} = \begin{cases} 1/2 & \text{if } n+1 > R, j > R \\ 1 & \text{if } n+1 > R, j \leq R \\ 0 & \text{if } n+1 < R \end{cases} \quad (6.11)$$

Writing the balance equations between the set of states $T_n = \{(n, j) : j = 0, \dots, S\}$ and $T_{n+1} = \{(n+1, j) : j = 0, \dots, S\}$, we have:

$$\sum_{j=0}^S \pi_{n,j} \mu = \sum_{j=0}^S \pi_{n+1,j} (\lambda_1 + \lambda_e q_{(n+1,j)}). \quad (6.12)$$

Following Equation (6.10), we can write

$$P_{n+1} \lambda_{(n+1)} = \sum_{j=0}^S \pi_{n+1,j} (\lambda_1 + \lambda_e q_{(n+1,j)}). \quad (6.13)$$

Therefore, from (6.12) and (6.13), we have the following result:

$$P_{n+1} \lambda_{(n+1)} = P_n \mu \quad (6.14)$$

On the other hand, using SQA we can model the network as $M_n/M/1$ which is a queue with conditional arrival rate $\lambda(n)$ and the steady state probability of having n items in the inventory x_n . So, writing the balance equation for this model will result in the following equation:

$$x_{n+1} \lambda_{(n+1)} = x_n \mu. \quad (6.15)$$

Comparing Equations (6.14) and (6.15) yield

$$x_n = P_n. \quad (6.16)$$

Namely, with the exact state-dependent arrival rates (λ_n) , the single queue approximation yields the same steady-state queue length distribution as in the original model of network. $\square$

6.1.4 Proof of Proposition 2.2

Based on the approximated value of λ_N in Equation (2.44), we can say that:

$$\frac{(1 - (1 - \Pr(I_c > R_c))^N)}{\Pr(I_c > R_c)} \geq 1. \quad (6.17)$$

Therefore:

$$(\lambda_e/N) \frac{(1 - (1 - \Pr(I_c > R_c))^N)}{\Pr(I_c > R_c)} \geq (\lambda_e/N). \quad (6.18)$$

This means that in the network based setting with the $\Pr(I_c > R_c)$ less than one, all producers will receive the second-class customers with a higher rate in compare to independent case. Decreasing the value of price p_e , the participants will try to allocate smaller part of their inventory for second-class demand, therefore the value of this probability will decrease and the difference between second-class customers' arrival rates in two cases becomes larger.

Therefore, using single queue approximation, we have two settings which differ in the rate of second-class customers' arrival. The first one is the independent setting with the profit function $\pi_{nc}(\lambda_1, (\lambda_e/N), S_{nc}^*, R_{nc}^*)$ and the second one is the cooperative case with the profit function $\pi_c(\lambda_1, (\lambda_e/N) + \Delta, S_c^*, R_c^*)$ with $0 \leq \Delta$.

Assuming that in the cooperative setting, we apply a control policy which filters the second-class customers and reduce the rate of arrival for second group of customers to (λ_e/N) and then use S_{nc}^* and R_{nc}^* as production and selling policies, we have:

$$\pi_c(\lambda_1, (\lambda_e/N) + \Delta, S_{nc}^*, R_{nc}^*) = \pi_{nc}^*(\lambda_1, (\lambda_e/N), S_{nc}^*, R_{nc}^*). \quad (6.19)$$

Also, we know that:

$$\{R_c^*, S_c^*\} = \operatorname{argmax}\{\pi_c(\lambda_1, (\lambda_e/N) + \Delta, S_c, R_c)\}. \quad (6.20)$$

Therefore:

$$\pi_c^*(\lambda_1, (\lambda_e/N) + \Delta, R_c^*, S_c^*) \geq \pi_c(\lambda_1, (\lambda_e/N) + \Delta, S_c, R_c). \quad (6.21)$$

Comparing Equations (6.19) and (6.21), we have:

$$\pi_c^*(\lambda_1, (\lambda_e/N) + \Delta, R_c^*, S_c^*) \geq \pi_{nc}^*(\lambda_1, (\lambda_e/N), R_{nc}^*, S_{nc}^*). \quad (6.22)$$

□

6.2 Appendix B

6.2.1 Proof of Proposition 3.1

Proof. We consider two systems, system $s1$ and $s2$ with different buffer capacities. System $s1$ has a buffer of capacity Z while system $s2$ has a capacity of $Z + \delta$, $\delta > 0$. The inputs used that are the time instances corresponding to the instances when the state of each station changes, and the maximum flow rates of each station at these time instances are kept the same for both systems. We use the notation $(k)_{s1}$ and $(k)_{s2}$ to denote the value of a variable k in system 1 and system 2 respectively.

For $t_i \in T^+$, tf_i in system $s1$ and $s2$ are determined as follows:

$$(tf_i)_{s1} = \frac{Z - (x(t_i))_{s1}}{u(t_i)} + t_i, \quad i \in I^+.$$

$$(tf_i)_{s2} = \frac{Z + \delta - (x(t_i))_{s2}}{u(t_i)} + t_i, \quad i \in I^+.$$

When both systems start at the same initial buffer level and the same transitions take place in both systems, we can show that $(x(t_i))_{s2} - (x(t_i))_{s1} \leq \delta$. That is the difference between the buffer trajectories cannot exceed the difference in the buffer capacity levels between these two systems. Consequently $(tf_i)_{s2} \geq (tf_i)_{s1}$ and Equation (3.5) yields

$$(\Delta p_i)_{s2} \geq (\Delta p_i)_{s1}.$$

Depending on whether system $s1$ and/or $s2$ reaches their respective buffer capacity level during $[t_i, t_{i+1})$, we have three different cases.

The first case is the situation in which both systems reach their buffer capacity levels in $[t_i, t_{i+1})$. Therefore,

$$t_{i+1} - t_i = (\Delta p_i)_{s1} + (\Delta f_i^-)_{s1}.$$

$$t_{i+1} - t_i = (\Delta p_i)_{s2} + (\Delta f_i^-)_{s2}.$$

The last three equations yield

$$(\Delta f_i^-)_{s1} \geq (\Delta f_i^-)_{s2}.$$

According to the Equation (3.31), the first machine will produce the material with rate $\mu_{\alpha_u(t_i)}^u(t_i)$ in Δp_i time duration and with rate $\mu_{\alpha_d(t_i)}^d(t_i)$ in Δf_i^- time duration if the second machine is in operational mode, i.e. $t_i \in T^+$ and $\mu_{\alpha_u(t_i)}^u(t_i) > \mu_{\alpha_d(t_i)}^d(t_i)$. It means that system $s2$ will produce the material with a greater rate $\mu_{\alpha_u(t_i)}^u(t_i)$ in a longer time duration $(\Delta p_i)_{s2}$ compared to system $s1$ which is producing with the same rate in a shorter time duration $(\Delta p_i)_{s1}$. As a result, we have $(P_u^+)_{s2} > (P_u^+)_{s1}$ for this case.

In the second case, the first system reaches its buffer capacity level, but the second one does not reach its buffer level. In this case,

$$t_{i+1} - t_i = (\Delta p_i)_{s1} + (\Delta f_i^-)_{s1}.$$

$$t_{i+1} - t_i = (\Delta p_i)_{s2}.$$

Using the same reasoning as before, we have more production in system $s2$ compared to the production in system $s1$ during $[t_i, t_{i+1})$

The third case is the situation in which both systems do not reach the buffer capacity level and we have $(\Delta p_i)_{s2} = (\Delta p_i)_{s1}$. Therefore both systems are producing with the same rate in the same interval of time so the amount of material they will produce will be the same $[t_i, t_{i+1})$.

As a result, we can conclude that:

$$(P_u^+)_{s2} \geq (P_u^+)_{s1}, \quad i \in I^+.$$

Now we consider $t_i \in T^-$ and examine Equation (3.32) for both systems. According to this equation, the first machine will produce with rate $\mu_{\alpha_u(t_i)}^u(t_i)$ in the whole time span between t_i and t_{i+1} and it does not depend on the time which takes for the system to reach zero buffer level. Therefore, we have:

$$(P_u^-)_{s2} = (P_u^-)_{s1}, \quad i \in I^-.$$

Finally, we consider $t_i \in T^0$. Equation (3.33) shows that the machine produces the material with rate $\mu_{\alpha_u(t_i)}^u(t_i)$ during the whole interval. As a result,

$$(P_u^0)_{s2} = (P_u^0)_{s1}, \quad i \in I^0.$$

Finally, Equation (3.34) gives

$$(PR)_{s2} \geq (PR)_{s1}.$$

□

6.2.2 Proof of Proposition 3.2

In order to transform the set of Equations (3.4)-(3.14) to a mathematical program with linear constraints, we defined the linear form of these equations as the constraints for the optimization problem by introducing binary variables in order to express equations that include comparisons between variables. The solution of the set of Equations (3.4)-(3.14) and the solution of the equations with linear constraints are equal to each other.

The feasible region of the mathematical program that includes the linear representation of the same set of equations will have a single point if and only if the solution of the set of Equations (3.4)-(3.14) is unique. In the case where the feasible solution is a single point, the objective function will not affect the solution.

Following this argument, we will prove the uniqueness of the solution for the iterative equations (3.4)-(3.14) in order to show the uniqueness of the solution for MILP. The proof is by induction:

When there is only one critical time, i.e. $N = 1$, we have $T = \{t_1\}$. If $t_1 \in T^+$,

$$tf_1 = \frac{Z - x(t_1)}{u(t_1)} + t_1,$$

$$\Delta p_1 = \min(t_2 - t_1, tf_1 - t_1),$$

$$\Delta f_1 = \max(0, t_2 - tf_1),$$

$$x(t_2) = x(t_1) + u(t_1)\Delta p_1.$$

Since $x(t_1)$ is given as the initial buffer level, the above equations yield a unique solution for tf_1 , Δp_1 , Δf_1 , and consequently for $x(t_2)$.

On the other hand, if $t_1 \in T^-$,

$$te_1 = \frac{-x(t_1)}{u(t_1)} + t_1,$$

$$\Delta e_1 = \min(t_2 - t_1, te_1 - t_1),$$

$$\Delta z_1 = \max(0, t_2 - te_1),$$

$$x(t_2) = x(t_1) + u(t_1)\Delta e_1.$$

Similarly, since $x(t_1)$ is given, the above equations yield a unique solution for te_1 , Δe_1 , Δz_1 , and then for $x(t_2)$.

Finally, if $t_1 \in T^0$,

$$x(t_2) = x(t_1).$$

So, when $x(t_1)$ is given $x(t_2)$ is determined uniquely.

Assuming that it is true for N critical times, i.e., $x(t_{N+1})$ is determined uniquely by $x(t_N)$, we now check whether when there are $N + 1$ critical times, $x(t_{N+2})$ is determined by $x(t_{N+1})$.

If $t_{N+1} \in T^+$,

$$tf_{N+1} = \frac{Z - x(t_{N+1})}{u(t_{N+1})} + t_{N+1},$$

$$\Delta p_{N+1} = \min(t_{N+2} - t_{N+1}, tf_{N+1} - t_{N+1}),$$

$$\Delta f_{N+1} = \max(0, t_{N+2} - tf_{N+1}),$$

$$x(t_{N+2}) = x(t_{N+1}) + u(t_{N+1})\Delta p_{N+1}.$$

Since $x(t_{N+1})$ is given, the above equations yield a unique solution for tf_{N+1} , Δp_{N+1} , Δf_{N+1} , and consequently for $x(t_{N+2})$.

If $t_{N+1} \in T^-$,

$$te_{N+1} = \frac{-x(t_{N+1})}{u(t_{N+1})} + t_{N+1},$$

$$\Delta e_{N+1} = \min(t_{N+2} - t_{N+1}, te_{N+1} - t_{N+1}),$$

$$\Delta z_{N+1} = \max(0, t_{N+2} - te_{N+1}),$$

$$x(t_{N+2}) = x(t_{N+1}) + u(t_{N+1})\Delta e_{N+1}.$$

Since $x(t_{N+1})$ is given, the above equations also yield a unique solution for te_{N+1} , Δe_{N+1} , Δz_{N+1} , and consequently for $x(t_{N+2})$.

Finally if $t_{N+1} \in T^0$,

$$x(t_{N+2}) = x(t_{N+1}).$$

As a result, when $x(t_1)$ is given, Equations (3.4)-(3.14) have a unique solution. The linear equations that represent the same equations have also a unique solution. As a result, there is only one feasible solution of the set of constraints in the MILP problem and this solution is the unique solution of the set of equations for all objective functions. Therefore, using MILP problem to simulate the system will give the same unique solution for system dynamics obtained by the iterative method. $\square$

6.3 Appendix c

6.3.1 Proof of Theorem 4.1

We use induction to prove the structural properties for all finite n . To start the induction, we set $v_0(x, 0) = 0$ and $v_0(x, 1) = 0$. For $n = 1$ all Inequalities (4.6), (4.7), (4.8), (4.9) and (4.10) are true. We assume that all these inequalities hold for n and prove them for $n + 1$. Hence,

1) For $n + 1$, we can write Inequality (4.6) as follows:

$$v_{n+1}(x, 1) - v_{n+1}(x + 1, 1) \leq v_{n+1}(x, 0) - v_{n+1}(x + 1, 0).$$

According to Inequalities (4.4) and (4.5), we have:

$$\begin{aligned} & \lambda_o \left[v_n(x+1, 1) - v_n(x+2, 1) \right] + \lambda_e \left[v_n(x, 1) - v_n(x+1, 1) \right] \\ & + \mu \left[\max \left\{ v_n(x-1, 0), v_n(x-1, 1) \right\} - \max \left\{ v_n(x, 0), v_n(x, 1) \right\} \right] \leq \\ & \lambda_o \left[\max \left\{ v_n(x+1, 0), v_n(x+1, 1) \right\} - \max \left\{ v_n(x+2, 0), v_n(x+2, 1) \right\} \right] \\ & + \lambda_e \left[v_n(x+1, 1) - v_n(x+2, 1) \right] + \mu \left[v_n(x, 0) - v_n(x+1, 0) \right]. \end{aligned} \quad (6.23)$$

In order to show that Inequality (6.23) is correct, we must show that the inequality holds for all the components associated with the same multipliers on the both sides.

For the multipliers of λ_o , we must show that:

$$v_n(x+1, 1) - v_n(x+2, 1) \leq \max \left\{ v_n(x+1, 0), v_n(x+1, 1) \right\} - \max \left\{ v_n(x+2, 0), v_n(x+2, 1) \right\}. \quad (6.24)$$

Case 1: If $x+1 \geq R$, then $\max \left\{ v_n(x+1, 0), v_n(x+1, 1) \right\} = v_n(x+1, 1)$ and $\max \left\{ v_n(x+2, 0), v_n(x+2, 1) \right\} = v_n(x+2, 1)$, so Inequality (6.24) is true.

Case 2: If $x+2 < R$, then $\max \left\{ v_n(x+1, 0), v_n(x+1, 1) \right\} = v_n(x+1, 0)$ and $\max \left\{ v_n(x+2, 0), v_n(x+2, 1) \right\} = v_n(x+2, 0)$, so Inequality (6.24) is true due to our assumption in Inequality (4.6).

Case 3: If $x+2 = R$, then $\max \left\{ v_n(x+1, 0), v_n(x+1, 1) \right\} = v_n(x+1, 0)$ and $\max \left\{ v_n(x+2, 0), v_n(x+2, 1) \right\} = v_n(x+2, 1)$, so Inequality (6.24) is true due to our assumption in Inequality (4.6).

For the multipliers of λ_e , we must show that:

$$v_n(x, 1) - v_n(x+1, 1) \leq v_n(x+1, 1) - v_n(x+2, 1). \quad (6.25)$$

Inequality (6.25) is true due to our assumption in Inequality (4.7) which implies that $v_n(x, 1)$ is concave.

For the multipliers of μ , we must show that:

$$\max \left\{ v_n(x-1, 0), v_n(x-1, 1) \right\} - \max \left\{ v_n(x, 0), v_n(x, 1) \right\} \leq v_n(x, 0) - v_n(x+1, 0). \quad (6.26)$$

Case 1: If $x < R$, then $\max \{v_n(x-1, 0), v_n(x-1, 1)\} = v_n(x-1, 0)$ and $\max \{v_n(x, 0), v_n(x, 1)\} = v_n(x, 0)$, so Inequality (6.26) is true due to our assumption in Inequality (4.8) which implies that $v_n(x, 0)$ is concave.

Case 2: If $x-1 > R$, then $\max \{v_n(x-1, 0), v_n(x-1, 1)\} = v_n(x-1, 1)$ and $\max \{v_n(x, 0), v_n(x, 1)\} = v_n(x, 1)$, so Inequality (6.26) is true due to our assumption in Inequalities (4.6) and (4.8).

Case 3: If $x = R$, then $\max \{v_n(x-1, 0), v_n(x-1, 1)\} = v_n(x-1, 0)$ and $\max \{v_n(x, 0), v_n(x, 1)\} = v_n(x, 1)$, so Inequality (6.26) is true due to our assumption in Inequality (4.8) and knowing that $v_n(x, 1) \geq v_n(x, 0)$.

2) For $n+1$, we can write Inequality 4.7 as follows:

$$v_n(x+2, 1) - v_n(x+1, 1) \leq v_n(x+1, 1) - v_n(x, 1)$$

and we have:

$$\begin{aligned} & \lambda_o [v_n(x+3, 1) - v_n(x+2, 1)] + \lambda_e [v_n(x+2, 1) - v_n(x+1, 1)] \quad (6.27) \\ & + \mu \left[\max \{v_n(x+1, 1), v_n(x+1, 0)\} - \max \{v_n(x, 0), v_n(x, 1)\} \right] \leq \\ & \lambda_o [v_n(x+2, 1) - v_n(x+1, 1)] + \lambda_e [v_n(x+1, 1) - v_n(x, 1)] \\ & + \mu \left[\max \{v_n(x, 1), v_n(x, 0)\} - \max \{v_n(x-1, 1), v_n(x-1, 0)\} \right] \end{aligned}$$

In order to show that Inequality (6.27) is correct, we must show that the inequality holds for all the components associated with the same multipliers on the both sides.

For the multipliers of λ_o , we must show that:

$$v_n(x+3, 1) - v_n(x+2, 1) \leq v_n(x+2, 1) - v_n(x+1, 1). \quad (6.28)$$

Inequality (6.28) is true due to our assumption in Inequality (4.7) which implies that $v_n(x, 1)$ is concave.

For the multipliers of λ_e , we must show that:

$$v_n(x+2, 1) - v_n(x+1, 1) \leq v_n(x+1, 1) - v_n(x, 1). \quad (6.29)$$

Inequality (6.29) is also true due to our assumption in Inequality (4.7).

For the multipliers of μ , we must show that:

$$\begin{aligned} & \max \left\{ v_n(x+1, 1), v_n(x+1, 0) \right\} - \max \left\{ v_n(x, 0), v_n(x, 1) \right\} \\ & \leq \max \left\{ v_n(x, 1), v_n(x, 0) \right\} - \max \left\{ v_n(x-1, 1), v_n(x-1, 0) \right\}. \end{aligned} \quad (6.30)$$

Case 1: If $x+1 < R$, then $\max \left\{ v_n(x+1, 1), v_n(x+1, 0) \right\} = v_n(x+1, 0)$, $\max \left\{ v_n(x, 0), v_n(x, 1) \right\} = v_n(x, 0)$ and $\max \left\{ v_n(x-1, 0), v_n(x-1, 1) \right\} = v_n(x-1, 0)$ so Inequality (6.30) is true due to our assumption in Inequality (4.8) which implies that $v_n(x, 0)$ is concave.

Case 2: If $x-1 > R$, then $\max \left\{ v_n(x+1, 1), v_n(x+1, 0) \right\} = v_n(x+1, 1)$, $\max \left\{ v_n(x, 0), v_n(x, 1) \right\} = v_n(x, 1)$ and $\max \left\{ v_n(x-1, 0), v_n(x-1, 1) \right\} = v_n(x-1, 1)$ so Inequality (6.30) is true due to our assumption in Inequality (4.7).

Case 3: If $x = R$, then $\max \left\{ v_n(x+1, 1), v_n(x+1, 0) \right\} = v_n(x+1, 1)$, $\max \left\{ v_n(x, 0), v_n(x, 1) \right\} = v_n(x, 1)$ and $\max \left\{ v_n(x-1, 0), v_n(x-1, 1) \right\} = v_n(x-1, 0)$. Therefore, we should show that the following inequality holds in this case:

$$v_n(x+1, 1) - v_n(x, 1) \leq v_n(x, 1) - v_n(x-1, 0). \quad (6.31)$$

According to the concavity conditions for $v_n(x, 0)$ and $v_n(x, 1)$, we can write:

$$v_n(x+1, 1) + v_n(x+1, 0) - v_n(x, 1) - v_n(x, 0) \leq v_n(x, 1) + v_n(x, 0) - v_n(x-1, 1) - v_n(x-1, 0). \quad (6.32)$$

On the other hand, we know that the following inequality is true because $v_n(x-1, 0) \geq v_n(x-1, 1)$.

$$v_n(x+1, 0) - v_n(x, 0) \leq v_n(x, 0) - v_n(x-1, 1). \quad (6.33)$$

From Inequalities (6.32) and (6.33) we can conclude that Inequality (6.31) is true because both $v_n(x, 0)$ and $v_n(x, 1)$ are decreasing functions.

The same argument also applies to show the concavity of $v_n(x, 0)$ and the same proof procedure can be used to show that Inequality (4.8) is true.

3) For $n + 1$, we can write Inequality (4.9) as follows:

$$\begin{aligned} \lambda_o \left[v_n(x+2, 1) + p_o \right] + \lambda_e v_n(x+1, 1) + \mu \max \left\{ v_n(x, 0), v_n(x, 1) \right\} - h(x) &\leq (6.34) \\ \lambda_o \left[v_n(x+1, 1) + p_o \right] + \lambda_e v_n(x, 1) + \mu \max \left\{ v_n(x-1, 0), v_n(x-1, 1) \right\} - h(x-1) & \end{aligned}$$

The inequalities in the first two lines can easily be shown to hold and the last line is trivially true. Therefore, $v_n(x, 1)$ is decreasing in x .

The same argument also applies to show the $v_n(x, 0)$ is decreasing in x . $\square$

6.3.2 Proof of Proposition 4.1

Based on the queueing model with threshold level policy R , we can obtain the steady state probabilities of the system as follows:

$$Pr(x_1 = l, x_2 = 0) = \left(\frac{1 - \rho}{1 + \lambda_e/\mu} \right) \left(\frac{\rho^{l-R}(\lambda_1 + \lambda_e)}{\mu} - \frac{\lambda_e \lambda_1^{l-R}}{(\mu + \lambda_1)^{l-R+1}} \right) \quad \forall l > R-1. \quad (6.35)$$

$$Pr(x_1 = l, x_2 = 1) = \left(\frac{1 - \rho}{1 + \lambda_e/\mu} \right) \left(\frac{\lambda_e \lambda_1^{l-R+1}}{(\mu + \lambda_1)^{l-R+2}} \right) \quad \forall l \geq R-1. \quad (6.36)$$

Therefore the value of $\bar{L}_Q$ is equal to:

$$\bar{L}_Q = \frac{(R-1)(\mu - \lambda_1) + (\lambda_1 + \lambda_e)(\mu(R-1) + \lambda_1(2-R))}{(\mu + \lambda_e)(\mu - \lambda_1)}. \quad (6.37)$$

According to Equation (6.37), we should set the value of R to its minimum possible value to maximize the long run profit for each member in the platform. Hence, the value of threshold level, R , is equal to one. $\square$

BIBLIOGRAPHY

- Akçay Y, Tan B (2008) On the benefits of assortment-based cooperation among independent producers. *Production and Operations Management* 17(6):626–640
- Alfieri A, Matta A (2012a) Mathematical programming formulations for approximate simulation of multistage production systems. *European Journal of Operational Research* 219(3):773–783
- Alfieri A, Matta A (2012b) Mathematical programming representation of pull controlled single-product serial manufacturing systems. *Journal of Intelligent Manufacturing* 23(1):23–35
- Altman E, Jimenez T, Koole G (1998) On optimal call admission control. In: *Decision and Control, 1998. Proceedings of the 37th IEEE Conference on*, vol 1, pp 569–574
- Anick D, Mitra D, Sondhi M (1982) Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal*, The 61(8):1871–1894
- Benjaafar S, ElHafsi M, Véricourt Fd (2004) Demand allocation in multiple-product, multiple-facility, make-to-stock systems. *Management Science* 50(10):1431–1448
- Benjaafar S, Cooper WL, Kim JS (2005) On the benefits of pooling in production-inventory systems. *Management Science* 51(4):548–565
- Benjaafar S, Li Y, Xu D, Elhedhli S (2008) Demand allocation in systems with multiple inventory locations and multiple demand sources. *Manufacturing & Service Operations Management* 10(1):43–60
- Bertsekas DP, Bertsekas DP, Bertsekas DP, Bertsekas DP (1995) *Dynamic programming and optimal control*, vol 1. Athena scientific Belmont, MA

- Bertsimas D (2014) Statistics and machine learning via a modern optimization lens. In: INFORMS The 2014-2015 Philip McCord Morse Lecture
- Bhandari A, Scheller-Wolf A, Harchol-Balter M (2008) An exact and efficient algorithm for the constrained dynamic operator staffing problem for call centers. *Management Science* 54(2):339–353
- Bierbooms JJPH (2012) Performance analysis of production lines: discrete and continuous flow models. PhD thesis, Technische Universiteit Eindhoven
- Binkhuysen FB, Ranschaert E (2011) Teleradiology: evolution and concepts. *European Journal of Radiology* 78(2):205–209
- Bispo C (2011) Single server scheduling problem. optimal policy for convex costs depends on arrival rates. In: Proc. of MISTA 2011-5th Multidisciplinary International Scheduling Conference
- Brouns GA, Van Der Wal J (2006) Optimal threshold policies in a two-class preemptive priority queue with admission and termination control. *Queueing Systems* 54(1):21–33
- Bucklin LP, Sengupta S (1993) Organizing successful co-marketing alliances. *The Journal of Marketing* 57(2):32–46
- Bynum AB, Irwin CA, Cranford CO, Denny GS (2003) The impact of telemedicine on patients' cost savings: some preliminary findings. *Telemedicine Journal and e-Health* 9(4):361–367
- Carr S, Duenyas I (2000) Optimal admission control and sequencing in a make-to-stock/make-to-order production system. *Operations research* 48(5):709–720
- Chang H (2015) Evaluation framework for telemedicine using the logical framework approach and a fishbone diagram. *Healthcare Informatics Research* 21(4):230–238

- Çil EB, Örmeci EL, Karaesmen F (2009) Effects of system parameters on the optimal policy structure in a class of queueing control problems. *Queueing Systems* 61(4):273–304
- Cohen MA, Kleindorfer PR, Lee HL (1988) Service constrained (s, S) inventory systems with priority demand classes and lost sales. *Management Science* 34(4):482–499
- Colledani M, Gershwin S (2013) A decomposition method for approximate evaluation of continuous flow multi-stage lines with general markovian machines. *Annals of Operations Research* 209(1):5–40
- Craig J, Petterson V (2005) Introduction to the practice of telemedicine. *Journal of Telemedicine and Telecare* 11(1):3–9
- Dallery Y (1994) On modeling failure and repair times in stochastic models of manufacturing systems using generalized exponential distributions. *Queueing Systems* 15(1-4):199–209
- Danilovic M, Winroth M (2005) A tentative framework for analyzing integration in collaborative manufacturing network settings: a case study. *Journal of Engineering and Technology Management* 22(1):141–158
- Deshpande V, Cohen MA, Donohue K (2003) A threshold inventory rationing policy for service-differentiated demand classes. *Management Science* 49(6):683–703
- Devaraj SJ, Ezra K (2011) Current trends and future challenges in wireless telemedicine system. In: *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*, vol 4, pp 417–421
- Diamantidis A, Papadopoulos C, Vidalis M (2004) Exact analysis of a discrete material three-station one-buffer merge system with unreliable machines. *International Journal of Production Research* 42(4):651–675
- Duran S, Liu T, Simchi-Levi D, Swann JL (2007) Optimal production and inventory policies of priority and price-differentiated customers. *IIE Transactions* 39(9):845–861

- Elwalid A, Mitra D (1991) Analysis and design of rate-based congestion control of high speed networks, i: stochastic fluid models, access regulation. *Queueing Systems* 9(1-2):29–63
- Erdogan SA, Krupski TL, Lobo JM (2018) Optimization of telemedicine appointments in rural areas. *Service Science* 10(3):261–276
- Ferreira J, Gigante F, Sarraipa J, Nunez MJ, Agostinho C, Jardim-Goncalves R (2014) Collaborative production using dynamic manufacturing networks for SME's. In: *Engineering, Technology and Innovation (ICE)*, 2014 International ICE Conference on, IEEE, pp 1–7
- Fleuren S, Bierbooms R, Adan I (2014) Performance analysis of exponential multi-server production lines with fluid flow and finite buffers. *Stochastic Models* 30(4):469–493
- Frank KC, Zhang RQ, Duenyas I (2003) Optimal policies for inventory systems with priority demand classes. *Operations Research* 51(6):993–1002
- Gayon JP, Benjaafar S, De Véricourt F (2009a) Using imperfect advance demand information in production-inventory systems with multiple customer classes. *Manufacturing & Service Operations Management* 11(1):128–143
- Gayon JP, De Vericourt F, Karaesmen F (2009b) Stock rationing in an M/Er/1 multi-class make-to-stock queue with backorders. *IIIE Transactions* 41(12):1096–1109
- Gershwin S, Berman O (1981) Analysis of transfer lines consisting of two unreliable machines with random processing times and finite storage buffers. *AIIE Transactions* 13(1):2–11
- Gershwin S, Werner L (2007) An approximate analytical method for evaluating the performance of closed-loop flow systems with unreliable machines and finite buffers. *International Journal of Production Research* 45(14):3085–3111
- Gupta V, Balter MH, Sigman K, Whitt W (2007) Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation* 64(9):1062–1081

- Gustke SS, Balch DC, West VL, Rogers LO (2000) Patient satisfaction with telemedicine. *Telemedicine Journal* 6(1):5–13
- Ha AY (1997a) Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science* 43(8):1093–1103
- Ha AY (1997b) Stock-rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Research Logistics* 44(5):457–472
- Hailey D, Ohinmaa A, Roine R (2004) Study quality and evidence of benefit in recent assessments of telemedicine. *Journal of Telemedicine and Telecare* 10(6):318–324
- Harrison JM (1975) Dynamic scheduling of a multiclass queue: Discount optimality. *Operations Research* 23(2):270–282
- Helber S, Jusić H (2004) A new decomposition approach for non-cyclic continuous material flow lines with a merging flow of material. *Annals of Operations Research* 125(1-4):117–139
- Helber S, Schimmelpfeng K, Stolletz R, Lagershausen S (2011) Using linear programming to analyze and optimize stochastic flow lines. *Annals of Operations Research* 182(1):193–211
- Hosseini B, Tan B (2017) Simulation and optimization of continuous-flow production systems with a finite buffer by using mathematical programming. *IIE Transactions* 49(3):255–267
- Hosseini B, Tan B (2019) Modeling and analysis of a cooperative production network. *International Journal of Production Research* (forthcoming)
- Huang B, Iravani SM (2007) Optimal production and rationing decisions in supply chains with information sharing. *Operations Research Letters* 35(5):669–676
- Hung K, Zhang YT (2003) Implementation of a wap-based telemedicine system for patient monitoring. *IEEE Transactions on Information Technology in Biomedicine* 7(2):101–107

- Iravani SM, Krishnamurthy V, Chao GH (2007) Optimal server scheduling in nonpreemptive finite-population queueing systems. *Queueing Systems* 55(2):95–105
- Ivanov D, Dolgui A, Sokolov B, Werner F, Ivanova M (2016) A dynamic model and an algorithm for short-term supply chain scheduling in the smart factory Industry 4.0. *International Journal of Production Research* 54(2):386–402
- Kim J, Gershwin S (2005) Integrated quality and quantity modeling of a production line. *OR spectrum* 27(2-3):287–314
- Koole G (1998) Structural results for the control of queueing systems using event-based dynamic programming. *Queueing Systems* 30(3-4):323–339
- Körpeoğlu E, Kurtz Z, Kılınç-Karzan F, Kekre S, Basu PA (2014) Business analytics assists transitioning traditional medicine to telemedicine at virtual radiologic. *Interfaces* 44(4):393–410
- Kouikoglou V, Phillis Y (1994) Discrete event modeling and optimization of unreliable production lines with random rates. *IEEE Transactions on Robotics and Automation* 10(2):153–159
- Ku CY, Jordan S (2003) Near optimal admission control for multiserver loss queues in series. *European Journal of Operational Research* 144(1):166–178
- LeBihan H, Dallery Y (2000) A robust decomposition method for the analysis of production lines with unreliable machines and finite buffers. *Annals of Operations Research* 93(1-4):265–297
- Lee JE, Hong Y (2003) A stock rationing policy in a (s, S) controlled stochastic production system with 2-phase coxian processing times and lost sales. *International Journal of Production Economics* 83(3):299–307
- Lemon C, Liu N, Lane S, Sud A, Branley J, Khadra M, Kim J (2018) Changes in user perceptions of a telemedicine system over time: from initial implementation to everyday use. *Telemedicine and e-Health* 24(7):552–559

- Levantese R, Matta A, Tolio T (2003) Performance evaluation of continuous production lines with machines having different processing times and multiple failure modes. *Performance Evaluation* 51(2):247–268
- Li BH, Zhang L, Wang SL, Tao F, Cao J, Jiang X, Song X, Chai X (2010) Cloud manufacturing: a new service-oriented networked manufacturing model. *Computer integrated manufacturing systems* 16(1):1–7
- Lippman SA, Ross SM (1971) The streetwalker’s dilemma: A job shop model. *SIAM Journal on Applied Mathematics* 20(3):336–342
- Liu S, Song M, Tan KC, Zhang C (2015) Multi-class dynamic inventory rationing with stochastic demands and backordering. *European Journal of Operational Research* 244(1):153–163
- Matta A (2008) Simulation optimization with mathematical programming representation of discrete event systems. In: *Proceedings of the 40th Conference on Winter Simulation, Winter Simulation Conference*, pp 1393–1400
- Mitra D (1988) Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Advances in Applied Probability* pp 646–676
- Nadar E, Akan M, Scheller-Wolf A (2016) Experimental results indicating lattice-dependent policies may be optimal for general assemble-to-order systems. *Production and Operations Management* 25(4):647–661
- Nahmias S, Demmy WS (1981) Operating characteristics of an inventory system with rationing. *Management Science* 27(11):1236–1245
- Niyato D, Hossain E, Diamond J (2007) Ieee 802.16/wimax-based broadband wireless access and its application for telemedicine/e-health services. *IEEE Wireless Communications* 14(1):72–83
- Norris AC, Norris A (2002) *Essentials of Telemedicine and Telecare*. Wiley Online Library

- Örmeci EL, Burnetas A (2004) Admission control with batch arrivals. *Operations Research Letters* 32(5):448–454
- Örmeci EL, van der Wal J (2006) Admission policies for a two class loss system with general interarrival times. *Stochastic Models* 22(1):37–53
- Örmeci EL, Burnetas A, van der Wal J (2001) Admission policies for a two class loss system. *Stochastic Models* 17(4):513–539
- Özdoğru Ü, Altıok T (2003) Analysis of two-valve fluid-flow systems with general repair times. In: *Analysis and modeling of manufacturing systems*, Springer, pp 255–288
- Pan X, Song J, Zhang B (2017) Patient admission control in a hierarchical healthcare system. *IFAC-PapersOnLine* 50(1):4636–4641
- Pang Z, Shen H, Cheng T (2014) Inventory rationing in a make-to-stock system with batch production and lost sales. *Production and Operations Management* 23(7):1243–1257
- Papakostas N, Efthymiou K, Georgoulas K, Chryssolouris G (2013) On the configuration and planning of dynamic manufacturing networks. In: *Robust Manufacturing Control*, Springer, pp 247–258
- Pawar P, Jones V, Van Beijnum BJF, Hermens H (2012) A framework for the comparison of mobile patient monitoring systems. *Journal of Biomedical Informatics* 45(3):544–556
- Peirce SC, Hardisty AR, Preece AD, Elwyn G (2011) Designing and implementing telemonitoring for early detection of deterioration in chronic disease: defining the requirements. *Health Informatics Journal* 17(3):173–190
- Pietrabissa A (2008) An alternative lp formulation of the admission control problem in multiclass networks. *IEEE Transactions on Automatic Control* 53(3):839–845

- Poffe A, Gershwin S (2005) Integrating quality and quantity modelling in a production line. Technical Report ORC-377-05, Massachusetts Institute of Technology Operations Research Center Working Paper Series
- Puterman ML (2014) Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons
- Putnik G (2012) Advanced manufacturing systems and enterprises: Cloud and ubiquitous manufacturing and an architecture. *Journal of Applied Engineering Science* 10(3):127–134
- Rahwan T, Michalak TP, Wooldridge M, Jennings NR (2015) Coalition structure generation: A survey. *Artificial Intelligence* 229:139–174
- Rajan B, Tezcan T, Seidmann A (2018) Service systems with heterogeneous customers: Investigating the effect of telemedicine on chronic care. *Management Science*, <https://doi.org/101287/mnsc20172979>
- Ren L, Zhang L, Wang L, Tao F, Chai X (2014) Cloud manufacturing: key characteristics and applications. *International Journal of Computer Integrated Manufacturing* (Published online):1–15
- Righter R (2000) Expulsion and scheduling control for multiclass queues with heterogeneous servers. *Queueing Systems* 34(1-4):289–300
- Roine R, Ohinmaa A, Hailey D (2001) Assessing telemedicine: a systematic review of the literature. *Canadian Medical Association Journal* 165(6):765–771
- Rykov V (2001) Monotone control of queueing systems with heterogeneous servers. *Queueing systems* 37(4):391–403
- Schruben LW (2000) Mathematical programming models of discrete event system dynamics. In: *Proceedings of the 32nd conference on Winter simulation*, Society for Computer Simulation International, pp 381–385

- Schwarz J, Stolletz R (2013) A sampling approach for the analysis of time-dependent stochastic flow lines. *Proceedings of the Ninth International Conference on Stochastic Models of Manufacturing and Service Operations* pp 181–188
- Scott RE, Mars M (2015) Telehealth in the developing world: current status and future prospects. *Smart Homecare Technology and TeleHealth* 3:25–37
- Sorknæs AD, Madsen H, Hallas J, Jest P, Hansen-Nord M (2011) Nurse teleconsultations with discharged copd patients reduce early readmissions—an interventional study. *The Clinical Respiratory Journal* 5(1):26–34
- Swaminathan V, Moorman C (2009) Marketing alliances, firm networks, and firm value creation. *Journal of Marketing* 73(5):52–69
- Tan B (2006) Modelling and analysis of a network organization for cooperation of manufacturers on production capacity. *Mathematical Problems in Engineering* 2006:1–24
- Tan B (2015) Mathematical programming representations of the dynamics of continuous-flow production systems. *IIE Transactions* 47(2):173–189
- Tan B, Akçay Y (2014) Assortment-based cooperation between two make-to-stock firms. *IIE Transactions* 46(3):213–229
- Tan B, Gershwin S (2009) Analysis of a general markovian two-stage continuous-flow production system with a finite buffer. *International Journal of Production Economics* 120(2):327–339
- Tan B, Gershwin S (2011) Modelling and analysis of markovian continuous flow systems with a finite buffer. *Annals of Operations Research* 182(1):5–30
- Tolio T, Matta A, Gershwin S (2002) Analysis of two-machine lines with multiple failure modes. *IIE Transactions* 34(1):51–62

- Topkis DM (1968) Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. *Management Science* 15(3):160–176
- Turhan A, Alanyali M, Starobinski D (2012) Optimal admission control in two-class preemptive loss systems. *Operations Research Letters* 40(6):510–515
- Ulukus MY, Güllü R, Örmeci EL (2011) Admission and termination control of a two class loss system. *Stochastic Models* 27(1):2–25
- Véricourt F, Karaesmen F, Dallery Y (2002) Optimal stock allocation for a capacitated supply system. *Management Science* 48(11):1486–1501
- Véricourt Fd, Karaesmen F, Dallery Y (2001) Assessing the benefits of different stock-allocation policies for a make-to-stock production system. *Manufacturing & Service Operations Management* 3(2):105–121
- Wadhwa S, Mishra M, Chan FT (2009) Organizing a virtual manufacturing enterprise: an analytic network process based approach for enterprise flexibility. *International Journal of Production Research* 47(1):163–186
- Warren I, Weerasinghe T, Maddison R, Wang Y (2011) Odintelehealth: A mobile service platform for telehealth. *Procedia Computer Science* 5:681–688
- Whitten P, Holtz B, Nguyen L (2010) Keys to a successful and sustainable telemedicine program. *International Journal of Technology Assessment In Health Care* 26(2):211–216
- Wijngaard J (1979) The effect of interstage buffer storage on the output of two unreliable production units in series, with different production rates. *AIIE transactions* 11(1):42–47
- Wootton R, Vladzimirsky A, Zolfo M, Bonnardot L (2011) Experience with low-cost telemedicine in three different settings. recommendations based on a proposed framework for network performance evaluation. *Global Health Action* 4(1):7214

- Wu D, Greer MJ, Rosen DW, Schaefer D (2013) Cloud manufacturing: Strategic vision and state-of-the-art. *Journal of Manufacturing Systems* 32(4):564–579
- Yang J, Qin Z (2007) Capacitated production control with virtual lateral transshipments. *Operations Research* 55(6):1104–1119
- Yeralan S, Tan B (1997) A station model for continuous materials flow production systems. *International Journal of Production Research* 35(9):2525–2542
- Yeralan S, Franck W, Quasem M (1986) A continuous materials flow production line model with station breakdown. *European Journal of Operational Research* 27(3):289–300
- Yu Y, Benjaafar S, Gerchak Y (2015) Capacity sharing and cost allocation among independent firms with congestion. *Production and Operations Management* 24(8):1285–1310
- Zhang L, Luo Y, Tao F, Li BH, Ren L, Zhang X, Guo H, Cheng Y, Hu A, Liu Y (2014) Cloud manufacturing: a new manufacturing paradigm. *Enterprise Information Systems* 8(2):167–187
- Zhao Gx, Wang Sy (2009) New research on a serving system under non-preemptive priority protocol. In: *International Conference on Information Engineering and Computer Science, 2009. ICIECS 2009.*, pp 1–4
- Zhao H, Deshpande V, Ryan JK (2006) Emergency transshipment in decentralized dealer networks: When to send and accept transshipment requests. *Naval Research Logistics (NRL)* 53(6):547–567
- Zhao H, Ryan JK, Deshpande V (2008) Optimal dynamic production and inventory transshipment policies for a two-location make-to-stock system. *Operations Research* 56(2):400–410