Response Time Reduction and Service-Level Differentiation in Supply Chain Design: Models and Solution Approaches

by

Navneet Vidyarthi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Management Sciences

Waterloo, Ontario, Canada, 2009

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

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

Abstract

Make-to-order (MTO) and assemble-to-order (ATO) systems are emerging business strategies in managing responsive supply chains, characterized by high product variety, highly variable customer demand, and short product life cycle. Motivated by the strategic importance of response time in today's global business environment, this thesis presents models and solution approaches for response time reduction and service-level differentiation in designing MTO and ATO supply chains.

In the first part, we consider the problem of response time reduction in the design of MTO supply chains. More specifically, we consider an MTO supply chain design model that seeks to simultaneously determine the optimal location and the capacity of distribution centers (DCs) and allocate stochastic customer demand to DCs, so as to minimize the response time in addition to the fixed cost of opening DCs and equipping them with sufficient assembly capacity and the variable cost of serving customers. The DCs are modelled as M/G/1 queues and response times are computed using steady-state waiting time results from queueing theory. The problem is set up as a network of spatially distributed M/G/1 queues and modelled as a nonlinear mixed-integer program. We linearize the model using a simple transformation and a piece-wise linear and concave approximation. We present two solution procedures: an exact solution approach based on cutting plane method and a Lagrangean heuristic for solving large instances of the problem. While the cutting plane approach provides the optimal solution for moderate instances in few iterations, the Lagrangean heuristic succeeds in finding feasible solutions for large instances that are within 5% from the optimal solution in reasonable computation times. We show that the solution procedure can be extended to systems with multiple customer classes. Using a computational study, we also show that substantial reduction in response times can be achieved with minimal increase in total costs in the design of responsive supply chains. Furthermore, we find the supply chain configuration (DC location, capacity, and demand allocation) that considers congestion and its effect on response time can be very different from the traditional configuration that ignores congestion.

The second part considers the problem of response time reduction in the design of a two-echelon ATO supply chain, where a set of plants and DCs are to be established to distribute a set of finished products with non-trivial bill-of-materials to a set of customers with stochastic demand. The model is formulated as a nonlinear mixed integer programming problem. Lagrangean relaxation exploits the echelon structure of the problem to decompose into two subproblems - one for the make-to-stock echelon and the other for the MTO echelon. We use the cutting plane based approach proposed above to solve the MTO echelon subproblem. While Lagrangean relaxation provides a lower bound, we present a heuristic that uses the solution of the subproblems to construct an overall feasible solution. Computational results reveal that the heuristic solution is on average within 6% from its optimal.

In the final part of the thesis, we consider the problem of demand allocation and capacity selection in the design of MTO supply chains for segmented markets with service-level differentiated customers. Demands from each customer class arrives according to an independent Poisson process and the customers are served from shared DCs with finite capacity and generally distributed service times. Service-levels of various customer classes are expressed as the fraction of their demand served within a specified response (sojourn) time. Our objective is to determine the optimal location and the capacity of DCs and the demand allocation so as to minimize the sum of the fixed cost of opening DCs and equipping them with sufficient capacity and the variable cost of serving customers subject to service-level constraints for multiple customer classes. The problem is set up as a network of spatially distributed M/M/1 priority queues and modelled as a nonlinear mixed integer program. Due to the lack of closed form solution for service-level constraints for multiple classes, we present an iterative simulation-based cutting plane approach that relies on discrete-event simulation for the estimation of the service-level function and its subgradients. The subgradients obtained from the simulation are used to generate cuts that are appended to the mixed integer programming model. We also present a near-exact matrix analytic procedure to validate the estimates of the service-level function and its subgradients from the simulation. Our computational study shows that the method is robust and provides an optimal solution in most of the cases in reasonable computation time. Furthermore, using computational study, we examine the impact of different parameters on the design of supply chains for segmented markets and provide some managerial insights.

Acknowledgements

My advisors
Prof. Samir Elhedhli & Prof. Elizabeth Jewkes
For their patience, tolerance, guidance, advice, enthusiasm, and support
For their dedication, selfless devotion, and enlightenment

My committee members
Prof. James Bookbinder, Prof. Miguel Anjos &
Prof. Anthony Vannelli
For their constructive comments, valuable suggestions, and referring
For their timely guidance and valuable advice

Prof. Dmitry Krass
Rotman School of Management, University of Toronto
For his insightful comments and suggestions

Department of Management Sciences & University of Waterloo
For the admission, ambience, faculty, staff, facilities, and financial support

Waterloo Management of Integrated Manufacturing Systems
(WATMIMS) Research Group
For the affiliation and facilities

Natural Sciences and Engineering Research Council of Canada &
Ontario Ministry of Training, Colleges, and Universities
For the financial support through research grants and scholarships

Geetali Dutta Vidyarthi & Sangeeta Vidyarthi
For their unconditional solidarity, endless love, support, forbearance, and patience

Other Members of Our Family
For their faith, encouragement, and non-intervention

Friends, Fellow Graduate Students, & Colleagues
For the quality time and the great support

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the early 1980s, most manufacturing and service organizations believed that low cost and high quality were the most fundamental sources of competitive advantage [79]. Increasingly stiff competitive pressure forced these organizations to continually search for strategies to design and produce higher quality products at more competitive prices. New manufacturing technologies and strategies, such as just-in-time manufacturing, kanban, lean manufacturing, and total quality management became very popular and organizations invested heavily in implementing these strategies. However, with the passage of time, as more and more manufacturers managed to reduce costs and improve quality, they realized that these technologies and strategies had already helped them reduce costs as much as is practically possible [127]. In the early 1990s, market competition further forced organizations to introduce new products and ever greater variety at rapid rates, and "speed" evolved as the competitive paradigm. In order to position themselves to succeed, organizations shifted their focus from *cost-based* and *quality-based* strategies to *speed-based* strategies. This new paradigm is termed as *time-based competition* [28, 135, 136].

With the emergence of time-based competition, *time (speed)* became the key competitive priority and one of the main drivers of supply chain's performance. Kevin Rollins, the vice chairman of Dell Computer Corporation, states "Most of the managerial challenges at Dell Computer have to do with what we call velocity - speeding the pace of every element of our business. Life cycles in our business are measured in month, not years, and if you don't move fast, you're out of the game" [89]. In today's business environment, an organization that has the ability to respond to customer demand within shorter time frames - both in terms of volume change and variety change, enjoys a competitive advantage. Organizations today compete on three basic temporal aspects: *product development cycle time*, *manufacturing lead time*, and *response time* [79]. Product development cycle time is the time needed to transform a design to a product. Manufacturing lead time is the time it takes to convert raw materials to finished products. Examples of response time include on-time delivery (a percentage of a match between the promised delivery date and the actual product delivery date, order processing time (time elapsed between the placement of order by a customer and the delivery of products to the customer), transit time (duration between the time of shipment and the time of receipt), and cash-to-cash cycle time (the amount of time required from the time a product has begun its manufacturing until the time it is completely sold. Shorter product development time gives an organization an early entry into the market, enabling it to establish itself as a market leader. Shorter manufacturing lead time allows the manufacturer to reduce finished goods inventories as well as work-in-process inventories, which in turn helps mitigate the risk of obsolescence and cut inventory costs. Shorter response time increases customer satisfaction which leads to a higher market share [28, 79, 135, 136]. Recent industry practice reveals that customers are even willing to pay a price premium for shorter response times [28].

Amongst all the three temporal aspects, *response time* is viewed as the external performance measure that represents the firm's commitment on customer satisfaction and it is this measure that the customer really cares. Therefore, many firms use response time standards as an explicitly advertised competitive instrument. Hence, the focus of this thesis is on response time reduction and response time based service-level differentiation in the design of supply chains.

1.1 Market-Responsive Supply Chain

For a variety of reasons, product and technology life cycles are shortening, competitive pressures force more frequent product changes and consumers demand a greater variety of products than ever before. The main challenge facing supply chains today is to quickly respond to frequent and unpredictable changes in the market [117]. Furthermore, many firms introduce innovations in trend, fashion, and technology to differentiate themselves from their competitor and to attract consumers to buy their offerings, which in turn increases their profit margins. Market-responsive supply chains have emerged as an ideal strategy to deal with such "innovative" products (e.g. laptops) whose demand is highly variable, product variety is high, life cycle is short, and market trends are changing.

Fisher [63] points out the difference between market-responsive supply chains and efficient supply chains. Efficient supply chains work well for 'functional" products (e.g. staple foods, toothpaste) manufactured using traditional mass production where there is low product variety, the cycle times are long and demand is fairly stable and predictable. While the primary focus of an efficient supply chain is to match supply to market demand at the lowest possible cost, a responsive supply chain focuses on responding quickly to unpredictable demand in order to minimize stockouts, forced markdowns, and obsolete inventory. The product design strategy of an efficient supply chain is to maximize performance and minimize cost, whereas a responsive chain uses modular design in order to postpone product differentiation for as long as possible. An efficient supply chain tends to explore ways to shorten lead time as long as it doesn't increase cost, whereas in a responsive supply chain, the focus is to invest aggressively in ways to reduce lead time.

Unlike efficient supply chains, a market-responsive supply chain deals with demand uncertainty using three strategies: (1) reduce uncertainty - for example by finding sources of new data that can serve as leading indicators or by having different products share common components/platform as much as possible so that demand for components becomes more predictable, (2) avoid uncertainty by reducing lead times and increasing supply chain flexibility so that the product can be made to order or assembled to order, and (3) hedge against uncertainty with buffers of excess capacity. Fisher [63] describes how companies such as National Bicycle and Sport Obermeyer illustrate different ways of blending these three strategies to create responsive supply chains. In this thesis, we attempt to incorporate these strategies in the strategic design of responsive supply chains.

1.2 Make-to-Order and Assemble-to-Order Supply Chains

Make-to-order (MTO) and assemble-to-order (ATO) systems are successful business strategies in managing market-responsive supply chains, characterized by *high product variety*, *highly variable customer demand*, and *short product life cycles*. Because of mass customization and competition on product variety, many firms adopt an MTO/ATO strategy to offer a variety of products and deal with product proliferation [137]. Dell's manufacturing and distribution of Personal Computers (PCs) is an excellent example of an MTO/ATO supply chain [50, 89]. Dell typically offers several lines of product, with each allowing at least dozens of "features" from which customers can select when placing an order - different combinations of CPU, hard drive, memory, and other peripherals. In Dell's supply chain, multiple components are procured and kept in inventory at various assembly facilities, from which they are assembled into a wide variety of finished products in response to customer orders. Whereas each of these components takes a substantial lead time to manufacture, the time to assemble all these components into a PC is low, provided there is sufficient assembly capacity and the components are available. In traditional make-to-stock (MTS) supply chains, the customer orders are met from stocks of an inventory of finished products that are kept at various points in the network. This is done to reduce the delay in fulfilling customer orders, increase sales, and avoid stockouts. However, the problems associated with holding inventory of finished products may outweigh the benefits, especially when those products become obsolete as technology advances or fashion changes. While an MTO strategy eliminates finished goods inventories and reduces a firm's exposure to the risk of obsolescence, it usually spells long response times (or order-to-delivery lead times) [69].

In order to reconcile the dual needs of short response time and high product variety, many firms such as General Motors, General Electric, American Standard, Compaq, IBM, BMW, Nike iD, and National Bicycle have adopted an ATO strategy. ATO is a hybrid strategy (i.e. mix of MTO and MTS) in which a subassembly, or a number of common subassemblies used in several products, are assembled and placed in inventory until the order is received for finished product [133]. This allows the firm to customize their orders by having the product ready using MTO strategy, while taking the advantage of economies of scale using MTS strategy. Also, the investment in semi-finished product inventory is smaller compared to the option of maintaining a similar amount of finished goods inventory. Furthermore, demand pooling benefits can be realized. Although maintaining a semi-finished product inventory in ATO systems lowers the customer response time as compared to a pure MTO system, it can further be reduced by minimizing congestion at the point of differentiation. Naturally, the response time to deliver the product is critical and forms the basis for competition. The potential for increased demand and/or consumers' willingness to pay a price premium for customized products within a shorter response time provides further incentives for firms to reduce response time in MTO and ATO supply chains.

1.3 Response Time Reduction

With the emergence of *speed* as a key competitive priority in the arena of global competition, numerous firms have reduced their response time. For example, Toyota reduced the lead time to deliver a custom built car from 30-60 days to within 5 days of receiving the customer order [128]. In 2000, General Motors announced a dramatic vision to reduce lead-time from 50-60 days to less than 10 days [127]. Ford planned to reduce the order-to-delivery cycle times to less than 15 days by 2000 [140]. By 2005, Nissan expected to cut the order-to-delivery cycle time from 40 days to 14 days initially, and then to 7-days, thereby saving $ 3600 per vehicle [115]. FAW-Volkswagen Automotive Company[1] (NE China) plans to reduce order-to-delivery cycle time from a few months to 3 weeks. In fashion retail sector, Liz Claiborne reduced the lead time from 10-50 weeks to less than 60 days by launching a campus in China [117]. Motorola delivers their customized cellular phones the next day to customers anywhere in the United States [137].

Furthermore, many firms use response-time standards as an advertising strategy in their promotion campaigns while others strive to reduce their response time in

[1] FAW-Volkswagen is a joint venture between First Automobile Works (FAW) and Volkswagen Group

an effort to improve their customer service-level. For example, Domino's offers a guarantee on the speed of its delivery. Under its "30 minutes or its free" policy, it offers the delivery free of charge if the customer order were to take more than 30 minutes. Citicorp introduced MortgagePower, which promised its customers a loan commitment within 15 days while others required 30 to 60 days to make a commitment [136]. Black Angus Restaurants offer their customers free lunches if not served within 10 minutes [132]. Banks like Wells Fargo markets its "five minute maximum wait policy", under which any customer having to wait more than five minutes in line receives five dollars [132]. Lucky, a supermarket chain in California, guarantees under its "3 is a crowd" marketing campaign that a new checkout counter will be opened if there are more than three customers waiting in line [132]. In freight services, Federal Express offers next day package delivery by 11:00 a.m. whereas UPS guarantees next day package delivery by 8:30 a.m. Furthermore, most major electronic brokerage firms, (e.g. Ameritrade, Fidelity, E-trade) all prominently feature the average or median execution speed per transaction which is monitored by independent firms. Some firms go as far as to provide an individual execution time score card as part of the customer's personal account statements. As an additional example, in the airline industry, independent government agencies (e.g. the Aviation Consumer Protection Division of the U.S. Department of Transportation, as well as internet travel services e.g Expedia) report the average delay on a flight by flight basis.

1.4 Service-Level Differentiation for Multiple Customer Segments

Service-level differentiation is a successful business strategy used by manufacturing and service firms in managing customer segments with different profitability and service quality requirements and expectations. Firms segment customers into multiple classes to which they offer the same product or service but with different levels of service quality so as to maximize (long run) profits. Customers may be segmented based on the price they pay, the volume they purchase, or the length of the contract they purchase. Examples of such market segmentation are abundant [7].

Computer software and hardware firms (e.g. IBM, Dell Inc.) often segment their customers into Home and Home Office users, Small Businesses, Large Businesses and the Government, and Education and Health Care sectors, where each segment is served according to a specific priority discipline based on their waiting time expectations. Airlines offer different check-in and security clearance procedures for their Economy, Business and First class passengers. Banks and credit card companies segment their customers into regular, Gold and Platinum customers. Amazon.com differentiates consumers based on their delivery time requirements by offering them to choose between expedited shipping or free shipping options.

Customers in distinct market segments view service quality differently. Service quality offered to customers may be specified in terms of penalties for delays, expediting undertaken in case of shortages, or guarantees of a fill rate for customers with long-term contracts. More specifically, in manufacturing sectors, typical measures of service levels include fill rate, expected order delay, the probability that the order delay does not exceed a quoted lead time, and the percentage of orders fulfilled accurately. In service industries, service levels are measured through expected customer waiting time, the probability that the customer receives service within a specified time window, or the probability that a customer does not leave before being served [20]. In MTO and ATO systems, where no finished product inventory is held in the system, the service levels are often specified as a function of response time (e.g. probability that the customer response time does not exceed a quoted lead time). The supply chain network must be designed to meet the service level requirements for multiple customer classes in such responsive supply chains.

1.5 Scope of This Research

The strategic importance of response time reduction and service-level differentiation in supply chains has made time a competitive priority. Such time-based competition is especially important in MTO and ATO supply chains, where multiple components or subassemblies are produced and kept in inventories, from which they can be rapidly assembled into a wide variety of finished products in response to customer orders. When products are made to order or assembled to order, there are no finished goods inventories to handle spikes in demand. Instead production or assembly capacity must deal with all the orders within an acceptable time frame, otherwise lost sales can easily wipe out firm's profit margin and diminish its competitive edge. Hence to reduce long response times or order-to-delivery lead-time, firms should invest heavily in acquiring sufficient production or assembly capacity [69]. Most of the models of MTO/ATO in the literature assume that the system has sufficient production/assembly capacity and the time to assemble a product from its components is negligible [133]. While this is a reasonable assumption for tactical and operational level planning, capacity decisions should be considered as one of the most critical decisions needed to realize a responsive supply chain during strategic planning and supply chain design. Transportation, inventory, production, and distribution decisions can often be changed frequently (or in the short term) in order to respond to the changes in factors such as the availability of raw materials, labor costs, transportation costs, inventory holding costs, exchange rates. Capacity decisions, on the other hand, are often fixed and difficult to change in the medium term as it involves huge investment. Therefore, insufficient capacity at production facilities and DCs may result in long response times and lost sales, no matter how well the transportation, inventory, production, and distribution decisions are optimized in response to changing conditions.

Chopra and Meindl [41] point out that response time in supply chains is influenced by the design of its distribution network. Firms that target customers who can tolerate long response times usually require few DC locations that may be far from the customers and can focus on increasing the capacity of the each location. On the other hand, firms that target customers who value short response time need to locate many facilities with adequate capacity to avoid congestion. Thus, a decrease in the response time customers desire increases the number and/or the size of the facilities in the network. However, supply chain network design models presented in the literature, mainly consider the design of "efficient" supply chains, where the focus is on cost reduction under a fairly stable and deterministic demand settings. Additionally, Lee and Billington [86] identify the separation of strategic supply chain design from operational decisions as one of the pitfalls of supply chain management. They state that when companies add or close a plant or distribution center in a supply chain network, the effect of changes in the network configuration on operational factors such as response time is often an afterthought. To the best of our knowledge, no research has modelled MTO and ATO supply chain network design problems from a response-time perspective. In this thesis, we consider the problem of designing such responsive supply chains that incorporate response time and demand uncertainty into the strategic design and planning. More specifically, we present models that capture the tradeoffs among response time costs, location and capacity acquisition costs, and transportation costs, while designing a one-echelon and two-echelon supply chain network. Given the fact the inclusion of response time makes the model intractable, we present efficient solution approaches that can solve such integrated models of supply chain network design.

1.6 Structure of the Thesis

The remainder of the thesis is organized as follows. Having outlined the motivation and the scope of this work, we proceed in Chapter 2 to describe the problem of response time reduction in MTO supply chain characterized by stochastic customer demand that has to be satisfied from a set of DCs, where sufficient capacity has to be acquired in order to avoid long response times. We briefly discuss the characteristics of the context in which the modelling is done. The variables and the parameters are established, the assumptions are stated, and a non-linear mixed integer programming model is formulated. We linearize the model using a simple transformation and piece-wise linear and concave approximations. We present two solution approaches: an exact solution procedure based on cutting plane method and a Lagrangean heuristic. Detailed computational results are reported. Through a case study, we show that substantial reduction in response time can be achieved with minimal increase in total costs in the design of responsive supply chains. Furthermore, we demonstrate the difference between the MTO supply chain configuration that considers congestion and its effect on response time and the traditional configuration that ignores congestion. We extend the basic model to multiple customer classes and general demand and service time distributions where DCs are modelled as spatially distributed GI/G/1 queues. Some of the results appear in Vidyarthi et al. [147].

In Chapter 3, we consider the problem of response time reduction in designing a two-echelon ATO supply chain. We present a model that locates plants and DCs, determines plant and DC capacity, and allocates customers to DCs with the objective of minimizing response time costs in addition to the fixed location and capacity acquisition costs, variables costs of transportation. We formulate the model as a nonlinear MIP problem and derive some valid cuts. Lagrangean relaxation is applied to decompose the problem by echelon resulting in two subproblems - one for MTS echelon and the other for MTO echelon. We use the solution methodology proposed above to solve the subproblem related to the MTO echelon to optimality. While Lagrangean relaxation provides a lower bound, we present a heuristic that uses the solution of the subproblems to construct an overall feasible solution. Detailed computational results are reported. The results appear in Vidyarthi et al. [147].

In Chapter 4, we consider the problem of designing an MTO supply chain for segmented markets with service-level differentiated customers. The modelling context is described, the assumptions are stated, and a model is formulated. Due to the lack of closed-form expressions for service-level constraints for multiple customer classes, we rely on discrete-event simulation for the estimation of the function and its subgradients. This subgradient information is used to generate constraints that are appended as cuts in an iterative cutting plane algorithm. We present a near-exact matrix analytic method to validate the estimates obtained from the simulation in some cases. We provide computational results and some managerial insights.

Finally, in Chapter 5, we summarize the contributions and outline some future research directions.

Chapter 2

Response Time Reduction in MTO Supply Chain Design[*]

2.1 Introduction

Although MTO systems are widely used business strategies in managing responsive supply chains, characterized by high product variety, highly variable customer demand, and short product life cycles, they usually spell long response time. Due to fierce pressure from global competition, this strategy needs to be supported by efficient design of a supply chain network that can meet the customer demand in reasonable response time. The objective of this chapter is to model the effect of response time reduction on supply chain network configuration and analyze the tradeoff among response time costs, facility location and capacity acquisition costs, and outbound transportation costs in the design of supply chain networks. More specifically, we present a model to determine the configuration of an MTO supply chain, where the emphasis is on minimizing customer response time through the acquisition of sufficient assembly capacity and the optimal allocation of workload to the assembly facilities (DCs) under stochastic customer demand settings. The DCs are modelled as spatially distributed queues with Poisson arrivals and general service times to capture the dynamics of response time. The model is formulated

[*] A version of the materials in this chapter has appeared in the paper: N. K. Vidyarthi, S. Elhedhli, and E. M. Jewkes (2009), Response time reduction in make-to-order and assemble-to-order supply chain design, *IIE Transactions*, 41(5), 448-466.

as a nonlinear mixed integer programming (MIP) problem and is linearized using piecewise linear functions. We present a cutting plane algorithm that provides the optimal solution to the problem. Furthermore, we present a Lagrangean relaxation heuristic procedure for solving large scale instances of such integrated models. Explicit consideration of congestion effects and their impact on response time in making location, capacity, and allocation decisions in supply chains distinguishes this work from most other supply chain design models.

The rest of the chapter is organized as follows. In Section 2.2, we briefly review the related literature. Section 2.3 provides a nonlinear MIP formulation of the MTO supply chain design problem. In Section 2.4, we linearize the model using a simple transformation and piecewise linear and concave approximations. An exact solution procedure based on cutting plane method is presented in Section 2.5, whereas Section 2.6 presents a Lagrangean heuristic. The simplifications resulting from assuming exponentially distributed service times (M/M/1 case) and deterministic service times (M/D/1 case) are explicitly described in Section 2.7. Computational results and managerial insights are reported in Section 2.8. In Section 2.9, we extend our model to include multiple customer classes and general demand processes and service time distributions. Finally, in Section 2.10, we offer concluding remarks.

2.2 Related Literature

The related literature can be categorized into three groups: (i) lead time reduction and capacity planning with congestion in supply chains, (ii) supply chain network design and (iii) stochastic location model with immobile servers

2.2.1 Lead Time Reduction and Capacity Planning with Congestion in Supply Chains

Since the seminal publications on time-based competition by Blackburn [28] and Stalk and Hout [136], there has been extensive research on lead time reduction in various domains of supply chains [41]. The focus is mainly on two components of lead time: *replenishment/supply lead time* and *delivery lead time*. Ray et al. [111] analyze the effects of investment in *replenishment lead time* reduction in a single-echelon MTS firm which faces a constant demand rate and replenishes its raw material inventory from a supplier using a continuous review (Q, r) inventory policy. In general, research on investments in lead time reduction in a stochastic (Q, r) inventory model setting can be categorized into two groups: (i) those assuming that variability of the lead time demand is due to demand rate variations while the lead time duration is fixed; and (ii) those assuming that variability of the lead time demand is due to the variability of the lead time duration, while the demand rate is constant. See Ray et al. [111] and references therein for more details along these lines of research. Glasserman and Wang [68] study the trade-offs between *delivery lead time* and inventory at a fixed fill rate. So and Song [132] develop a model to study the interaction among *delivery lead time*, price, and capacity selection decisions. Their model simultaneously determines the optimal price, delivery lead time, and capacity decisions for a service facility modelled as an M/M/1 queue to maximize the overall profit. They assume that the demand is sensitive to both the price and delivery lead time and is modelled using a log-linear (Cobb-Douglas) demand function. Ray and Jewkes [112] study the interaction between *delivery lead time* and operating/capacity costs in an MTO environment when demand and price are lead time sensitive. While most of these works have dealt with replenishment or delivery lead times, the focus of our work is the reduction of *production lead time* in an MTO/ATO supply chain comprising of spatially distributed facilities, where the production lead time is determined by the demand allocated, capacity acquired at the facilities, and the variability in demand and processing time. We model facilities as M/G/1 queues and use steady state waiting time results from queueing theory to model production lead time.

On the other hand, *capacity planning and expansion problems*, which primarily deal with determining the locations, sizes, timings, as well as the types of production facilities to meet the demand at minimum total cost, have been widely studied in operations research/management literature for the last five decades [88]. However, the relationships between the capacity, demand, and operational performance measures such as lead times and congestion have only been addressed recently (see Kim and Uzsoy [80], Rajagopalan and Yu [108] and references therein). Most of these models have focused on a single-facility system with homogenous servers, fixed capacity, and a single customer class. The congestion is usually modelled as a constraint that ensures that a target lead time is satisfied with a pre-specified probability. Despite that, solution approaches proposed to date are either approximate or heuristic. Note that capacity planning of manufacturing facilities or DCs is a key issue in determining the order-to-delivery lead time especially in MTO supply chains, where customer orders arrive randomly and lead to high variability and congestion. The problem is further complicated by the presence of spatially distributed facilities with heterogenous servers that interact to satisfy the demand arising from multiple customer classes with varying order-to-delivery lead time expectations. In this thesis, we attempt to bridge this gap.

2.2.2 Supply Chain Network Design

Motivated by the importance and financial impact of strategic planning decisions in supply chains, numerous researchers have developed models and solution approaches for designing supply chain networks. These models cover formulations which range in complexity from simply *linear, single-echelon, single-commodity, single-period, single-country, uncapacitated, deterministic models* to *non-linear, multi-echelon, multi-commodity, multi-period*, *multi-national*, *capacitated*, *stochastic models*. Deterministic models for supply chain design assume that the parameters of the model are known with certainty, whereas stochastic models consider uncertainties in the environment by associating some probability distribution with the parameters. Successful applications of SCND models range from automotive, chemical, consumer goods, electronics, foods, to packaging industries. Firms such as Caterpillar [110], DowBrands [76], Volkswagen of America [77], Kellogg [31], Hewlett-Packard [84], BMW [64], Proctor & Gamble [34], Digital Equipment Corporation [12], Nabisco Bakeries [32], Ault Foods [106], Libbey-Owens-Ford [92], Hunt-Wesson-Foods [67] and several others have achieved substantial efficiency gains and cost reduction through the optimization of their supply chain network design. For an extensive review on supply chain design, readers are referred to Vidal and Geotschalckx [143], Erengüç et al. [59], Sarmiento and Nagi [116], Daskin et al. [46], Klose and Drexl [81], Martel [91], Meixell and Gargeya [93], Snyder and Daskin [129], and Shen [121]. As pointed out in the recent reviews, over the years, supply chain network design models have evolved to consider issues such as:

- Multinational or global logistics issues such as tarrifs, transfer prices, taxes [36, 42, 82]; NAFTA [113, 153]; exchange rates [144].
- Capacity acquisition and technology selection [45, 58, 104, 142].
- Inventory management [15, 16, 35, 44, 49, 60, 74, 95, 96, 100, 101, 102, 103, 114, 122, 123, 118, 119, 124, 120, 121, 125, 129, 134, 138, 139, 144, 145]; safety stock and risk pooling [134, 144, 145]
- Lead time [61, 62, 134, 144]
- Congestion [61, 62, 75, 114]
- Transportation mode selection [43, 62, 144]
- Facility reliability and disruption [130, 131], supplier reliability [107, 144]
- Robustness and risk-management [97]
- Bill-of-materials [104]
- Vehicle routing [30]
- Service-level measures such as fill rate [35], demand coverage [123]

Although various integrated models of supply chain design have been proposed in recent years to support lead time reduction, these models have continued to be largely guided by more traditional concerns of efficiency and cost in MTS settings, where the primary focus is on minimizing fixed cost of facility location and variable transportation cost under a fairly stable and deterministic customer demand settings. Some of them include Dogan and Goetschalckx [53], Vidal and Goetschalckx [144], Teo and Shu [139], Shen [118], Eskigun et al. [61, 62] and Elhedhli and Gzara [58]. For example, Vidal and Goetschalckx [144] present a model that captures the effect of change in transportation lead time and demand on the optimal configuration of the global supply chain network, assuming that the demand is deterministic. In their model, the transportation lead time is captured through safety stock to be held to meet the demand during stochastic replenishment lead times, which in turn depends on the transportation mode selected. Eskigun et al. [61, 62] incorporate delivery lead time and the choice of transportation mode in the design of supply chain under a deterministic demand setting. Their delivery lead time is composed of load make-up time (batching delays), queueing time in the facilities due to congestion, and unavoidable administrative delays in the system and is modelled using a simple function. Sourirajan et al. [134] extend their model to incorporate safety stock decisions to capture the tradeoffs between risk pooling and congestion as a result of consolidation of operations. These models tend to ignore congestion at the facilities and its effect on response time. Their solutions prescribe locating facilities whose capacity utilization is very high, resulting in excessively long response time when subjected to the variability in service times and randomness in customer orders. Reviews by Vidal and Goetschalckx [143], Erengüç [59], and Sarmiento and Nagi [116] also point out that most of the existing supply chain design models do not consider measures of customer service such as response time in making location/allocation decisions. Also refer to the recent review by Klose and Drexl [81]. This is not surprising given the complexity of the model and the interplay of locational and queueing aspects of the problem. To the best of our knowledge, Huang et al. [75] is one of the first papers to model the effect of congestion in the design of distribution networks. They model capacity using the mean and variance of the DCs as continuous variables, whereas our model considers capacity as a set of discrete options with known means and variances. They propose solution procedures based on outer approximation and Lagrangean relaxation, and test these on small instances of the problem.

2.2.3 Stochastic Location Model with Immobile Servers

Another growing body of literature that is related to our work and accounts for congestion in strategic planning are *stochastic location models with immobile servers* (SLMIS). SLMIS seeks to locate a set of service facilities with adequate capacity, and allocate stochastic demand to each of them, so as to minimize the fixed costs of opening facilities and acquiring service capacity, as well as the variable access and expected waiting time [57]. The problem arises in several planning contexts: location of emergency service facilities - such as medical clinics; police stations, and fire stations; refuse collection and disposal [5]; location of stores and service centers; telecommunication network design [40]; location of bank branches and automated teller machines [2, 148]; internet mirror site location [148]; and preventive healthcare facilities [155]. For an extensive review, the readers are referred to Berman and Krass [25] and Boffey et al. [29]. The study of models of this type originated with Berman et al. [27]. For further discussion of this class of problems, we refer the reader to Berman et al. [26], Marianov and Serra [90], Wang et al. [149], Elhedhli [57], Baron et al. [17] and references therein. As pointed out by Elhedhli [57], there are two different approaches in the literature to model service quality in stochastic location models. The first considers a probability constraint that ensures that waiting time or queue length does not exceed a certain threshold [90], whereas the second approach incorporates the service cost directly in the objective function [3, 5, 8, 37, 57, 148]. Due to the complexity of the underlying problem, most papers in this area make strong assumptions: Either the number/capacity of the facilities (or both) are assumed to be fixed or the facilities are uncapacitated [2, 3]. The demand arrival is assumed to be a (time homogeneous) Poisson process and the service process is assumed to be exponential [2, 3, 8, 9, 10, 24, 57, 90, 148]. Customers travel to the closest facility to obtain service [2, 3, 148]. Customers from different priority classes are often grouped into a single class [2, 3, 8, 9, 10, 24, 57, 90, 148]. In some situations (e.g. bank branches, ATMs, medical clinics), when the facilities offer a comparable level of service and little is known about the waiting times, then it is natural for the customer to select a nearest facility. However, in MTO/ATO supply chains, this is not the case. Furthermore, most of the previous work with the exception of Elhedhli [57] has been done in the context of coverage models, where the idea is to provide "adequate" service to the maximum number of customers. To the best of our knowledge, Baron et al. [17] is one of the few to consider the problem under much more general spatial distribution of demand arrival and service process, without fixing either the number or the capacity of the facilities.

Despite the aforementioned assumptions, most of the techniques proposed to date to solve these problems, with the exception of Elhedhli [57], are either approximate or heuristic based. For example, solution approaches developed include asymptotic approximation [37], Lagrangean relaxation based heuristics [5, 8, 9, 10, 148]; Greedy heuristic [149, 155]; Descent algorithm [2]; Ascent algorithm [23]; Simulated annealing [3, 23, 24]; Tabu search [148, 149]; Genetic Algorithm [3, 23, 24].

This is not surprising, given the complexity of the model due to nonlinearity in the objective function or the constraints, and the interplay of locational and queueing aspects of the problem.

2.3 Model Formulation

Consider the problem of designing an MTO supply chain, where a set of DCs are to be established and equipped with sufficient capacity to serve a set of customers (Figure 2.1). Sufficient capacity here implies being able to obtain service without waiting for an excessively long time after the order is placed. The DCs maintain an inventory of multiple components and facilitate the assembly and shipment of a wide variety of finished products in a timely fashion without carrying expensive finished-goods inventory and incurring long response time. In MTO supply chains, where customer orders triggers the assembly of finished product from components, response time consists of assembly lead time and delivery lead time. The delivery time between individual DCs and customers can be assumed to be relatively constant compared to the order fulfilment time at DCs in such settings. Moreover, it can further be reduced (using alternative transportation modes or expedited delivery services) to respond quickly to customer orders on a short term basis. However, assembly lead time is highly dependent on the DC capacity and the allocated workload and is difficult to change (on a short term basis) once the DC is established.

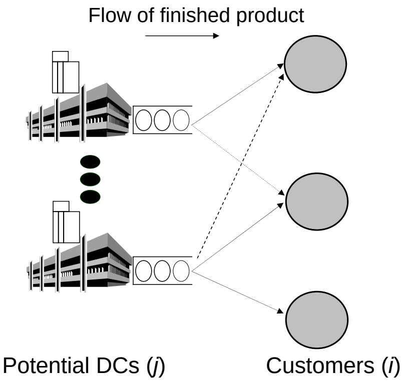


Figure 2.1: A make-to-order supply chain network

To model this, we assume that the demand for each product from each customer is independent and occurs according to a Poisson process. Once the demand for a product is realized at the customers' end, the order is placed at the DCs. DCs will act as assembly facilities and the customers' orders arriving at the DCs are met on a first-come first-serve (FCFS) basis. We assume that each DC operates as a single flexible-capacity server with an infinite buffer to accommodate customer orders waiting for service. We also assume that there is an unlimited supply of components and that inventory holding costs at the DCs are insignificant. Under these assumptions, the MTO supply chain is modelled as a network of independent M/G/1 queues in which the DCs are treated as servers with service rates proportional to their capacity. As pointed out by Baron et al. [17], there are two common ways to model flexible capacity of a queuing system. One is to assume multiple parallel servers each with a given service rate, and the other is to assume a single server with a flexible service rate. We model each DC as a single server with discrete capacity levels. The response time is computed using the expressions for steady state expected waiting time from queueing theory. As pointed out by Elhedhli [56] and Boffey et al. [29], there are two different approaches to incorporate response time in the model. The first incorporates response time cost in the objective function, whereas the second approach considers a probability constraint that ensures that waiting time or queue length does not exceed a certain threshold [17]. In this model, we use the first approach, where the response time cost is incorporated into the objective function. Later in Chapter 4, we describe a model with probability constraint that ensures that waiting time does not exceed a certain threshold.

Hence, the model formulated below simultaneously determines the location and capacity of DCs and the assignment of customer to DCs by minimizing the response time costs in addition to the fixed location and capacity acquisition costs, the assembly and transportation costs from DCs to customers. Besides capacity restrictions (steady state conditions) at the DCs, and the demand requirements, there are constraints which ensure that at most one capacity level is selected at the DCs. To model this problem, we define the following notation:

Indices and parameters:

i : Index for customers, $i = 1, 2, \ldots, I$.

j : Index for potential DCs, $j = 1, 2, \ldots, J$.

k : Index for potential capacity level at DCs, $k = 1, 2, \ldots, K$.

f_{jk} : Fixed cost of opening DC j and acquiring capacity level k (\$/period).

c_{ij} : Unit cost of serving customer i from DC j (\$/unit).

t : Mean response time cost per unit time per customer (\$/period/customer).

λ_i : Mean demand rate for the product from customer i (units/period).

μ_{jk} : Mean *service rate* at DC j, if it is allocated capacity level k (units/period).

σ^2_{jk} : Variance of *service times* at DC j, if it is allocated capacity level k.

Decision variables:

x_{ij} : Fraction of customer i's demand served by DC j $(0 \leq x_{ij} \leq 1)$.

y_{jk} : 1, if DC j is opened and capacity level k is acquired; 0, otherwise.

Let the demand for the product at customer location i be an independent random variable that follows a Poisson process with mean λ_i. If x_{ij} is the fraction of customer i's demand served by DC j, then the aggregate demand arrival rate at DC j is also a random variable that follows a Poisson process with mean $\lambda_j = \sum_{i=1}^{I} \lambda_i x_{ij}$, due to the superposition of Poisson processes. If the service times at each DC follows a general distribution and each DC is modelled as an M/G/1 queue, then the mean service rate of DC j, if it is allocated capacity level k, is given by $\mu_j = \sum_{k=1}^{K} \mu_{jk} y_{jk}$ and the variance in service times is $\sigma^2_j = \sum_{k=1}^{K} \sigma^2_{jk} y_{jk}$. This service rate reflects the server capacity or essentially the number of make-to-order products a DC can assemble and ship in a given time period. Let τ_j represent the mean service time at DC j $(\tau_j = 1/\mu_j)$, ρ_j be the utilization of DC j $(\rho_j = \lambda_j/\mu_j)$, and CV^2_j be the squared coefficient of variation of service times $(CV^2_j = \sigma^2_j/\tau^2_j)$. Under steady state conditions $(\lambda_j < \mu_j)$ and FCFS queuing discipline, the expected *average* waiting time (including the service time) at DC j is given by the Pollaczek-Khintchine (PK) formula:

$$\mathbb{E}[W_j(M/G/1)] = \left(\frac{1 + CV^2_j}{2}\right) \frac{\tau_j \rho_j}{1 - \rho_j} + \tau_j = \left(\frac{1 + CV^2_j}{2}\right) \frac{\lambda_j}{\mu_j(\mu_j - \lambda_j)} + \frac{1}{\mu_j} \quad \forall j$$

and the expected *total* waiting time for DC j is obtained by multiplying the waiting time at DC j by the expected demand as:

$$\left(\frac{1+CV_j^2}{2}\right)\frac{\lambda_j^2}{\mu_j(\mu_j-\lambda_j)}+\frac{\lambda_j}{\mu_j} \qquad \forall j$$

The expected *total* waiting time for the entire system is given by:

$$\begin{aligned}\mathbb{E}[W(M/G/1)] &= \sum_{j=1}^{J}\left[\left(\frac{1+CV_j^2}{2}\right)\frac{\lambda_j^2}{\mu_j(\mu_j-\lambda_j)}+\frac{\lambda_j}{\mu_j}\right]\\ &= \frac{1}{2}\sum_{j=1}^{J}\left[\left(1+CV_j^2\right)\frac{\lambda_j}{\mu_j-\lambda_j}+\left(1-CV_j^2\right)\frac{\lambda_j}{\mu_j}\right]\end{aligned}$$

This is equivalent to

$$\frac{1}{2}\sum_{j=1}^{J}\left\{\left(1+\sum_{k=1}^{K}CV_{jk}^2 y_{jk}\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk}y_{jk}-\sum_{i=1}^{I}\lambda_i x_{ij}}+\left(1-\sum_{k=1}^{K}CV_{jk}^2 y_{jk}\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk}y_{jk}}\right\} \tag{2.1}$$

The resulting non-linear MIP formulation is:

$$[P_N]:\min \quad \sum_{j=1}^{J}\sum_{k=1}^{K}f_{jk}y_{jk}+\sum_{i=1}^{I}\sum_{j=1}^{J}c_{ij}\lambda_i x_{ij}+t\mathbb{E}[W(M/G/1)] \tag{2.2}$$

$$\text{s.t.} \quad \sum_{i=1}^{I}\lambda_i x_{ij}\le\sum_{k=1}^{K}\mu_{jk}y_{jk} \qquad \forall j \tag{2.3}$$

$$\sum_{k=1}^{K}y_{jk}\le 1 \qquad \forall j \tag{2.4}$$

$$\sum_{j=1}^{J}x_{ij}=1 \qquad \forall i \tag{2.5}$$

$$0\le x_{ij}\le 1, \quad y_{jk}\in\{0,1\} \qquad \forall i,j,k \tag{2.6}$$

The first term in the objective function (2.2) represents the fixed cost of locating DCs and equipping them with adequate assembly capacity. Although the capital location cost is strategic and one-time, it is amortized over the span of the planning period. The second term accounts for the variable cost of assembly and shipment of products from DCs to customers. The third term is the *expected total response time cost* or the lost sales due to excessive response times between DCs and customers. The expected total response time cost is expressed as a product of average response time cost per unit time that one of its customers spends in the system and expected total waiting time in the system. It can be interpreted as a penalty function that reflects the true cost of not fulfilling customer orders in the committed lead time. In practice, determining the values of average response time cost can be challenging, however one can rely on techniques outlined in Rao et al. [110]. To accurately reflect lost sales due to unacceptable response time, Rao et al. [110] surveyed Caterpillar dealers to determine the percent of customers who would renege if a product was not available immediately, after two-weeks, and after four-weeks. A lower bound on the response time cost can be provided by the inventory holding costs [61]. Furthermore, the problems can be solved iteratively with different values of t to obtain a tradeoff curve from which decision makers may choose a solution based on their preference between location and capacity acquisition cost, transportation cost, and response time costs. In this chapter, for simplicity, we assume that the cost of response time is linearly proportional to the waiting time. However, the model can be extended to other cost functions such as a piecewise linear function or a cost function proportional to $\max\{0, \mathbb{E}[W(M/G/1)] - W_0\}$, where W_0 is the maximum tolerated waiting time for a customer. Moreover, the average response time cost t may vary from customer to customer (t_{ij}) if desired, but we assume for simplicity that it is the same across customers.

Constraints (2.3) ensure that the steady state conditions ($\lambda_j \leq \mu_j$) at the DCs are met. Constraint set (2.4) ensures that at most one capacity level is selected at a DC, whereas constraint set (2.5) ensures that the total demand is met. Constraints (2.6) are nonnegativity and binary restrictions. The formulation can easily handle single sourcing requirements by imposing binary restrictions on x_{ij}. This would restrict the assignment of customer i's demand to one and only one DC j.

The nonlinearity in $[P_N]$ arises due to the expression of the total waiting time at the DCs, $\mathbb{E}[W(M/G/1)]$. The results on the convexity of waiting time in a M/G/1 queue (with FCFS service discipline) by Tu and Kumin [141] and Weber [151] are worthwhile noting. Tu and Kumin [141] and Weber [151] have proved that in a M/G/1 queue, the expected steady-state waiting time is a nonincreasing convex function of the service rate and a nondecreasing convex function of the arrival rate. It can be shown that $\mathbb{E}[W(M/G/1)]$ is convex in aggregate arrival rate λ_j, for fixed value of μ_j and convex in service rate μ_j, for fixed value of λ_j, where $\lambda_j = \sum_{i=1}^{I} \lambda_i x_{ij}$ and $\mu_j = \sum_{k=1}^{K} \mu_{jk} y_{jk}$ as it satisfies the following properties: $\frac{\partial^2 W(\lambda_j,\mu_j)}{\partial \lambda^2} > 0$, $\frac{\partial^2 W(\lambda_j,\mu_j)}{\partial \mu^2} > 0$, and $\frac{\partial^2 W(\lambda_j,\mu_j)}{\partial \mu \partial \lambda} < 0$. Intuitively, one would expect that the waiting time increases with increasing marginal returns as the arrival rate increases and decreases with decreasing marginal returns as the service rate increases. In the next section, we deal with the nonlinearity due to the expression of total average waiting time using a linearization based on a simple transformation and a piecewise linear approximation. We also present an exact solution procedure based on the cutting plane method. Our cutting plane method is similar to the outer approximation method [54].

2.4 Linearization of the Model

In order to linearize equation (2.1), let us define nonnegative auxiliary variables R_j, such that

$$R_j = \frac{\lambda_j}{\lambda_j - \mu_j} = \frac{\sum_{i=1}^{I} \lambda_i x_{ij}}{\sum_{k=1}^{K} \mu_{jk} y_{jk} - \sum_{i=1}^{I} \lambda_i x_{ij}} \qquad \forall j$$

which implies that

$$\begin{aligned} \sum_{i=1}^{I} \lambda_i x_{ij} &= \frac{R_j}{1+R_j} \sum_{k=1}^{K} \mu_{jk} y_{jk} && (2.7) \\ &= \rho_j \sum_{k=1}^{K} \mu_{jk} y_{jk} = \sum_{k=1}^{K} \mu_{jk} z_{jk}, \quad \text{where} \quad z_{jk} = \begin{cases} 0 & \text{if} \quad y_{jk} = 0 \\ \rho_j & \text{if} \quad y_{jk} = 1 \end{cases} \quad \forall j, k && (2.8) \end{aligned}$$

and $\rho_j = \frac{R_j}{1+R_j}$ is the server (DC) utilization. Since there is at most one k' with $y_{jk'} = 1$ while $y_{jk} = 0$ for all other $k \neq k'$, the expression $z_{jk} = \rho_j y_{jk}$ can be ensured by adding the following constraints

$$z_{jk} \leq y_{jk} \qquad \forall j, k$$

$$\sum_{k=1}^{K} z_{jk} = \rho_j \qquad \forall j$$

The function $\rho_j = \frac{R_j}{1+R_j}$ is concave. Given a set of points R_j^h indexed by H, ρ_j can be approximated by an infinite set of piecewise linear functions that are tangent to ρ_j at points R_j^h (as shown in Figure 3.2) i.e.

$$\rho_j = \min_{h \in H} \left\{ \frac{1}{(1+R_j^h)^2} R_j + \frac{(R_j^h)^2}{(1+R_j^h)^2} \right\}, \qquad \forall j$$

$$\text{or} \quad \rho_j \leq \frac{1}{(1+R_j^h)^2} R_j + \frac{(R_j^h)^2}{(1+R_j^h)^2}, \qquad \forall j, h \in H$$

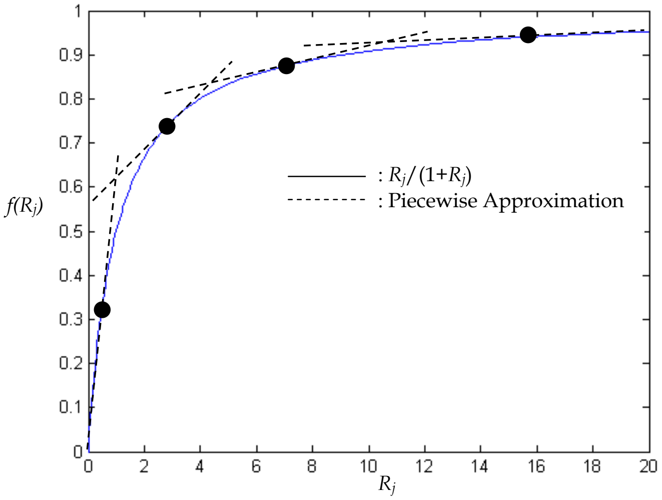


Figure 2.2: A piecewise linear approximation of $\frac{R_j}{1+R_j}$

The expression for $\mathbb{E}[W(M/G/1)]$ can be re-written as:

$$\begin{aligned}\mathbb{E}[W(M/G/1)] =& \frac{1}{2} \sum_{j=1}^{J} \left\{ \left(1 + \sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right) R_j + \left(1 - \sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right) \rho_j \right\} \\ =& \frac{1}{2} \sum_{j=1}^{J} \left(R_j + \sum_{k=1}^{K} CV_{jk}^2 w_{jk} + \rho_j - \sum_{k=1}^{K} CV_{jk}^2 z_{jk} \right)\end{aligned}$$

$$\text{where} \quad w_{jk} = \begin{cases} 0 & \text{if} \quad y_{jk} = 0 \\ R_j & \text{if} \quad y_{jk} = 1 \end{cases} \quad \text{and} \quad z_{jk} = \begin{cases} 0 & \text{if} \quad y_{jk} = 0 \\ \rho_j & \text{if} \quad y_{jk} = 1 \end{cases} \qquad \forall j, k$$

Similarly, because there exists at most one k' with $y_{jk'} = 1$ while $y_{jk} = 0$ for all other $k \neq k'$, the expression $w_{jk} = R_j y_{jk}$ can be ensured by adding the following constraints

$$\begin{aligned} & w_{jk} \leq M y_{jk} && \forall j, k \\ & \sum_{k=1}^{K} w_{jk} = R_j && \forall j \end{aligned}$$

where M is the usual Big-M.

The resulting linear MIP formulation is:

$$[P_{L(H)}] : \min \quad \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} y_{jk} + \sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}\lambda_i x_{ij} + \frac{t}{2}\sum_{j=1}^{J}\left\{R_j + \rho_j + \sum_{k=1}^{K} CV_{jk}^2 (w_{jk} - z_{jk})\right\} \tag{2.9}$$

$$\begin{aligned}
\text{s.t.} \quad & \sum_{i=1}^{I} \lambda_i x_{ij} - \sum_{k=1}^{K} \mu_{jk} z_{jk} = 0 && \forall j && (2.10) \\
& \sum_{k=1}^{K} y_{jk} \leq 1 && \forall j && (2.11) \\
& \sum_{j=1}^{J} x_{ij} = 1 && \forall i && (2.12) \\
& z_{jk} - y_{jk} \leq 0 && \forall j, k && (2.13) \\
& \rho_j - \frac{1}{(1+R_j^h)^2} R_j \leq \frac{(R_j^h)^2}{(1+R_j^h)^2} && \forall j, h \in H && (2.14) \\
& \rho_j - \sum_{k=1}^{K} z_{jk} = 0 && \forall j && (2.15) \\
& w_{jk} - M y_{jk} \leq 0 && \forall j, k && (2.16) \\
& \sum_{k=1}^{K} w_{jk} - R_j = 0 && \forall j && (2.17) \\
& y_{jk} \in \{0,1\}; \quad 0 \leq x_{ij}, z_{jk} \leq 1; \quad \rho_j, R_j, w_{jk} \geq 0; && \forall i, j, k && (2.18)
\end{aligned}$$

The steady state conditions ($\lambda_j < \mu_j$) translate into capacity constraints, and are enforced by the constraints (2.10) and (2.29) and forced to "<" by the term R_j in the objective.

2.5 Exact Solution Procedure

Note that $[P_{L(H)}]$ is a minimization problem, hence at least one of the constraints in (2.30) will be binding. This implies that

$$\rho_j = \min_{h \in H} \left\{ \frac{1}{(1+R_j^h)^2} R_j + \frac{(R_j^h)^2}{(1+R_j^h)^2} \right\} \qquad \forall j \quad \text{when} \quad y_{jk} = 1$$

In order to deal with the infinite number of constraints (2.30) in the linear MIP model $[P_{L(H)}]$, we use a *cutting plane method*, described as follows. For an initial and finite set of points $(R_j^h)_{\bar{H} \subset H}$, $[P_{L(\bar{H})}]$ is a relaxation of the full problem $[P_{L(H)}]$, hence a lower bound to $[P_{L(H)}]$ or $[P_N]$ is provided by the optimal objective function value $v(P_{L(\bar{H})})$, where

$$v(P_{L(\bar{H})}) = \sum_{j=1}^{J} \sum_{k=1}^{K} f_{jk} \overline{y}_{jk} + \sum_{i=1}^{I} \sum_{j=1}^{J} c_{ij} \lambda_i \overline{x}_{ij} + \frac{t}{2} \sum_{j=1}^{J} \left\{ \overline{R}_j + \overline{\rho}_j + \sum_{k=1}^{K} CV_{jk}^2 (\overline{w}_{jk} - \overline{z}_{jk}) \right\}$$

where $(\overline{x}, \overline{y}, \overline{R}, \overline{\rho}, \overline{w}, \overline{z})$ is the solution of $[P_{L(\bar{H})}]$.

Furthermore, $(\overline{x}, \overline{y})$ is feasible to $[P_N]$ and hence:

$$\sum_{j=1}^{J} \sum_{k=1}^{K} f_{jk} \overline{y}_{jk} + \sum_{i=1}^{I} \sum_{j=1}^{J} c_{ij} \lambda_i \overline{x}_{ij} +$$
$$\frac{t}{2} \sum_{j=1}^{J} \left\{ \left(1 + \sum_{k=1}^{K} CV_{jk}^2 \overline{y}_{jk} \right) \frac{\sum_{i=1}^{I} \lambda_i \overline{x}_{ij}}{\sum_{k=1}^{K} \mu_{jk} \overline{y}_{jk} - \sum_{i=1}^{I} \lambda_i \overline{x}_{ij}} + \left(1 - \sum_{k=1}^{K} CV_{jk}^2 \overline{y}_{jk} \right) \frac{\sum_{i=1}^{I} \lambda_i \overline{x}_{ij}}{\sum_{k=1}^{K} \mu_{jk} \overline{y}_{jk}} \right\}$$

provides an upper bound to $[P_{L(\bar{H})}]$ and $[P_N]$. If the best known upper bound coincides with the lower bound at a given iteration, then the optimal solution is obtained and the method is terminated. If not, a new set of cuts (3.14) are generated using $(\overline{R}_j)$ and appended to $[P_{L(\bar{H})}]$ and the procedure is repeated. The computational performance of the method is reported in Section 2.7.

2.6 Lagrangean Relaxation

Initial computational testing reveals that the direct solution of $[P_{L(\bar{H})}]$ takes long runtimes as the problem size increases. In this section, we propose a Lagrangean heuristic. Relaxing constraints (2.5) in $[P_N]$ with dual multipliers $\alpha_i, i = 1, ..., I$ leads to a Lagrangean nonlinear subproblem which decomposes into J independent subproblems, one for each DC j, as follows:

$$
\begin{aligned}
&[SP_{N(j,\alpha)}]: \\
&\min \quad \sum_{k=1}^{K} f_{jk} y_{jk} + \sum_{i=1}^{I} \left(c_{ij}\lambda_i - \alpha_i\right) x_{ij} + \\
&\qquad \frac{t}{2}\left(1 + \sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right) \frac{\sum_{i=1}^{I} \lambda_i x_{ij}}{\sum_{k=1}^{K} \mu_{jk} y_{jk} - \sum_{i=1}^{I} \lambda_i x_{ij}} + \\
&\qquad \frac{t}{2}\left(1 - \sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right) \frac{\sum_{i=1}^{I} \lambda_i x_{ij}}{\sum_{k=1}^{K} \mu_{jk} y_{jk}} \\
&\text{s.t.} \quad \sum_{i=1}^{I} \lambda_i x_{ij} \le \sum_{k=1}^{K} \mu_{jk} y_{jk} \\
&\qquad \sum_{k=1}^{K} y_{jk} \le 1 \\
&\qquad 0 \le x_{ij} \le 1, \quad y_{jk} \in \{0,1\} \qquad \forall i, k
\end{aligned}
$$

Given that at most one capacity level can be selected at DC j, the solution to subproblem $[SP_{N(j,\alpha)}]$ for DC j for a given set of multipliers α_i, can be obtained by solving K independent subproblems, one for each capacity level k:

$$
\begin{aligned}
[SP_{N(j,k,\alpha)}]: \min \quad & f_{jk} + \sum_{i=1}^{I} \left(c_{ij}\lambda_i - \alpha_i\right) x_{ij}^k + \\
& \frac{t}{2}\left\{ \left(1 + CV_{jk}^2\right) \frac{\sum_{i=1}^{I} \lambda_i x_{ij}^k}{\mu_{jk} - \sum_{i=1}^{I} \lambda_i x_{ij}^k} + \left(\frac{1 - CV_{jk}^2}{\mu_{jk}}\right) \sum_{i=1}^{I} \lambda_i x_{ij}^k \right\} \\
\text{s.t.} \quad & \sum_{i=1}^{I} \lambda_i x_{ij}^k \le \mu_{jk} \\
& 0 \le x_{ij}^k \le 1 \qquad \forall i
\end{aligned}
$$

We introduce the superscript k with the variable x_{ij} as it depends on μ_{jk}. Note that $[SP_{N(j,k,\alpha)}]$ is *continuous nonlinear knapsack problem.* Introducing auxiliary variables R_{jk}, where $R_{jk} = \sum_{i=1}^{I} \lambda_i x_{ij}^k / (\mu_{jk} - \sum_{i=1}^{I} \lambda_i x_{ij}^k)$, and using the linearization scheme proposed in section 3.3, problem $[SP_{N(j,k,\alpha)}]$ reduces to following linear program:

$$
\begin{aligned}
[SP_{L(j,k,\alpha,H)}] : \min \quad & f_{jk} + \sum_{i=1}^{I} (c_{ij}\lambda_i - \alpha_i)\, x_{ij}^k + \frac{t(1+CV_{jk}^2)}{2} R_{jk} + \frac{t(1-CV_{jk}^2)}{2\mu_{jk}} \sum_{i=1}^{I} \lambda_i x_{ij}^k \\
\text{s.t.} \quad & \sum_{i=1}^{I} \lambda_i x_{ij} \le \mu_{jk} \\
& \sum_{i=1}^{I} \lambda_i x_{ij}^k - \frac{\mu_{jk}}{(1+R_{jk}^h)^2} R_{jk} \le \frac{\mu_{jk}(R_{jk}^h)^2}{(1+R_{jk}^h)^2} \qquad \forall h \in H \\
& 0 \le x_{ij}^k \le 1, \quad R_{jk} \ge 0 \qquad \forall i, j
\end{aligned}
$$

which is amenable to solution by the cutting plane method of section 3.4. Note that the capacity constraints will never be binding at optimality, otherwise the R_{jk} term in the objective function goes to infinity. Furthermore, for a given set of dual multipliers, α_{ij} and DC j, the subproblem $[SP_{L(j,k,\alpha,H)}]$ is solved to optimality for each capacity level $k = 1, ..., K$. The level k^* that yields the most negative objective function value is assigned to DC j, and the variable y_{jk^*} is set to 1, whereas y_{jk}'s are set to zero, $\forall k \neq k^*$. If no such level k^* exists, then no DC is opened at location j, and the variables x_{ij} and y_{jk} are all set to zero.

2.6.1 The Lower Bound

The lower bound to $v(P_N)$ is given by $\sum_{j=1}^{J} \sum_{k=1}^{K} v(SP_{N(j,k,\alpha)}) + \sum_{i=1}^{I} \alpha_i$, where $v(\cdot)$ denotes the optimal objective value of the problem $(\cdot)$. The best lower bound is the solution of the Lagrangean dual problem,

$$v(LD) = \max_\alpha \left\{ \sum_{j=1}^{J} \sum_{k=1}^{K} v(SP_{N(j,k,\alpha)}) + \sum_{i=1}^{I} \alpha_i \right\},$$

which is equivalent to

$$
\max_\alpha \sum_{i=1}^{I} \alpha_i + \sum_{j=1}^{J} \min_{k,q \in Q_x^j} \left\{ 0, f_{jk} + \sum_{i=1}^{I} (c_{ij}\lambda_i - \alpha_i)\, x_{ij}^{kq} + \frac{t(1+CV_{jk}^2) \sum_{i=1}^{I} \lambda_i x_{ij}^{kq}}{2(\mu_{jk} - \sum_{i=1}^{I} \lambda_i x_{ij}^{kq})} + \frac{t(1-CV_{jk}^2)}{2\mu_{jk}} \sum_{i=1}^{I} \lambda_i x_{ij}^{kq} \right\}
$$

where Q_x^j is the index set of extreme point of the set $\{x : \sum_{i=1}^{I} \lambda_i x_{ij}^k \leq \mu_{jk}, \quad 0 \leq x_{ij}^k \leq 1\}$. The Lagrangean dual problem can be reformulated as a linear program with exponential number of constraints as:

$$
\begin{aligned}
[MP] : \max_{\alpha, \theta_j \leq 0} \quad & \sum_{i=1}^{I}\sum_{l=1}^{L} \alpha_{il} + \sum_{j=1}^{J} \theta_j \\
\text{s.t.} \quad & \theta_j + \sum_{i=1}^{I} x_{ij}^{kq} \alpha_i \leq f_{jk} + \sum_{i=1}^{I} c_{ij}\lambda_i x_{ij}^{kq} + \\
& \frac{t(1+CV_{jk}^2)}{2}\left(\frac{\sum_{i=1}^{I}\lambda_i x_{ij}^{kq}}{\mu_{jk} - \sum_{i=1}^{I}\lambda_i x_{ij}^{kq}}\right) + \frac{t(1-CV_{jk}^2)}{2\mu_{jk}}\sum_{i=1}^{I}\lambda_i x_{ij}^{kq} \\
& \qquad \forall j, k, q \in Q_x^0
\end{aligned}
$$

The relaxation of $[MP]$ defined on subsets $\bar{Q}_x^j \subset Q_x^j$ results in a relaxed master problem $[RMP]$. We use Kelley's classical cutting plane method [78], in which the optimal $\bar{\alpha}$ from $[RMP]$ is used to solve the subproblems $[SP_{N(j,k,\alpha)}]$ and generate $J \times K$ cuts of the form:

$$
\begin{aligned}
\theta_j + \sum_{i=1}^{I} \bar{x}_{ij}^{kq}\alpha_i \quad & \leq f_{jk} + \sum_{i=1}^{I} c_{ij}\lambda_i \bar{x}_{ij}^{kq} + \frac{t(1+CV_{jk}^2)}{2}\left(\frac{\sum_{i=1}^{I}\lambda_i x_{ij}^{kq}}{\mu_{jk} - \sum_{i=1}^{I}\lambda_i x_{ij}^{kq}}\right) + \\
& \frac{t(1-CV_{jk}^2)}{2\mu_{jk}}\sum_{i=1}^{I}\lambda_i x_{ij}^{kq}
\end{aligned}
$$

The index set $\bar{Q}_x^j$ is updated as iterations proceed. The Lagrangean process terminates when the gap between the objective of the master problem and subproblems is less than some pre-specified optimality tolerance ϵ specified by the user.

2.6.2 The Heuristic: Finding a Feasible Solution

The solution of the subproblems $[SP_{L(j,k,\alpha,H)}]$ provide the location of DCs, their capacity levels (y_{jk}), the allocation of customers to DCs (x_{ij}), and a lower bound to $[P_N]$. This solution may be infeasible for $[P_N]$ as the allocation of customers to DCs may not satisfy the demand assignment constraints $\sum_{j=1}^{J} x_{ij} = 1$, $\forall i$. To find an upper bound, we present a heuristic that attempts to construct a feasible solution at the final iteration of the Lagrangean dual problem. We fix the location and the capacity levels of the DCs for which $y_{jk} = 1$ at the final iteration of the Lagrangean dual problem. We compute the total service rate, $\sum_{j=1}^{J} \sum_{k=1}^{K} \mu_{jk} y_{jk}$ and the total arrival rate $\sum_{i=1}^{I} \lambda_i$. If the total service rate is less than the total arrival rate, then we pick a candidate DC (in order of its utilization ρ_j) and increase its capacity to the next level (i.e. set $y_{j,k+1} = 1$) if there exists one to choose from else we pick the next DC from the list until the constraint $\sum_{j=1}^{J} \sum_{k=1}^{K} \mu_{jk} y_{jk} \geq \sum_{i=1}^{I} \lambda_i$ is satisfied. In order to determine the optimal allocation of the customer demand to the open DCs, we solve the following LP:

$$\begin{aligned}
\min \quad & \sum_{i=1}^{I} \sum_{j=1}^{J} \left(c_{ij} + \frac{t(1 - CV_{jk}^2)}{2\mu_{jk}} \right) \lambda_i x_{ij} + \frac{t(1 + CV_{jk}^2)}{2} \sum_{j=1}^{J} \sum_{k=1}^{K} R_{jk} \\
\text{s.t.} \quad & \sum_{i=1}^{I} \lambda_i x_{ij} \leq \mu_{jk} \qquad \forall j \\
& \sum_{j=1}^{J} x_{ij} = 1 \qquad \forall i \\
& \sum_{i=1}^{I} \lambda_i x_{ij} - \frac{\mu_{jk}}{(1 + R_{jk}^h)^2} R_{jk} \leq \frac{\mu_{jk} (R_{jk}^h)^2}{(1 + R_{jk}^h)^2} \qquad \forall h \in H \\
& 0 \leq x_{ij} \leq 1, \quad R_{jk} \geq 0 \qquad \forall i, j
\end{aligned}$$

2.7 Special Cases

In this section, we look at the two special cases that are commonly looked at in the literature. In particular, we consider systems with exponential and deterministic service time distributions.

Case I. Systems with Exponential Service Times (M/M/1 case): The exponential processing and assembly time is a reasonable assumption in cases where there is high variability in setup times and processing times, e.g. in semiconductor wafer fabrication [79]. Also, this is more reasonable than deterministic processing and assembly times for MTO products with very high product variety and varying batch sizes. For exponentially distributed service times at the DCs, the total expected waiting time for the entire system is given by:

$$\mathbb{E}[W(M/M/1)] = \sum_{j=1}^{J} \frac{\lambda_j}{\mu_j - \lambda_j} = \sum_{j=1}^{J} \left(\frac{\sum_{i=1}^{I} \lambda_i x_{ij}}{\sum_{k=1}^{K} \mu_{jk} y_{jk} - \sum_{i=1}^{I} \lambda_i x_{ij}} \right) = \sum_{j=1}^{J} R_j$$

The resulting linear MIP model is:

$$\begin{aligned} [P_{L(H)}^{M/M/1}] : \min \quad & \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} y_{jk} + \sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}\lambda_i x_{ij} + t\sum_{j=1}^{J} R_j \qquad (2.19) \\ \text{s.t.} \quad & (10) - (15) \\ & y_{jk} \in \{0,1\}; \quad 0 \le x_{ij}, z_{jk} \le 1; \quad \rho_j, R_j \ge 0 \qquad \forall i,j,k \end{aligned}$$

The model is structurally identical to the service system design model presented in Amiri [8] and Elhedhli [57].

Case II. Systems with Deterministic Service Times (M/D/1 case): In many cases, the processing/assembly of finished products at the DCs often involves repeated steps without much variation [79]. This is particularly true for MTO products with limited options and batch size of one such as Dell PCs. For deterministic service times, the expected waiting time for the entire system is given by:

$$\mathbb{E}[W(M/D/1)] = \frac{1}{2}\sum_{j=1}^{J} \left(\frac{\lambda_j}{\mu_j - \lambda_j} + \frac{\lambda_j}{\mu_j} \right) = \frac{1}{2}\sum_{j=1}^{J} (R_j + \rho_j)$$

The resulting linear MIP model is as follows:

$$\begin{aligned} [P_{L(H)}^{M/D/1}] : \min \quad & \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} y_{jk} + \sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}\lambda_i x_{ij} + \frac{t}{2}\sum_{j=1}^{J} (R_j + \rho_j) \qquad (2.20) \\ \text{s.t.} \quad & (10) - (15) \\ & y_{jk} \in \{0,1\}; \quad 0 \le x_{ij}, z_{jk} \le 1; \quad \rho_j, R_j \ge 0 \qquad \forall i,j,k \end{aligned}$$

The cutting plane method described above can be used to solve these models to optimality. Alternatively, one can rely on the Lagrangean heuristic proposed above. Some computational results are provided in Section 2.8.

2.8 Computational Results and Insights

In this section, we report our computational experiences with the proposed solution methodologies and present some insights. All the proposed solution procedures were coded in C and the MIP problems were solved using ILOG CPLEX 10.1 (using the Callable Library) on a Sun Blade 2500 workstation with 1.6-GHz UltraSPARC IIIi processors. In the implementation of the iterative cutting plane method and after the solution of the relaxed MIP, we use the procedure $CPXaddrows()$ to append the cuts generated and exploit warm starting.

2.8.1 Test Problems

The test problems are derived from the 2000 census data consisting of 150 largest cities in the continental United States (see Daskin[47]). We generate nine sets of test problems by setting the number of customers (I) to the 50, 100, and 150 largest cities, and the potential DC locations (J) to the 5, 10, and 20 most populated cities. The mean customer demand rates λ_i are obtained by dividing the population of those cities by 10^3. The unit transportation costs c_{ij} are obtained by dividing the great-circle distance between the customer i and the potential DC location j by 100. The service rate of DC j equipped with capacity level k, is set to $\mu_{jk} = \beta_k \sum_i \lambda_i$ (where $\beta_k = 0.15, 0.20, 0.45$ for $I = 50$; $\beta_k = 0.10, 0.20, 0.30$ for $I = 100$; $\beta_k = 0.10, 0.15, 0.20, 0.30, 0.45$ for $I = 150$). The fixed costs, f_{jk} are set to $100 \times \sqrt{\mu}_{jk}$ to reflect economies of scale. For the M/M/1 case, the variance of service times σ^2_{jk} are set to the mean service rates, μ_{jk} whereas for the M/G/1, the σ^2_{jk} is obtained by setting coefficient of variation (CV) to 1.5. The average response time cost t is set to $\theta \times \frac{\sum_i \sum_j (\lambda_i c_{ij})}{I \times J}$, where θ is the response time cost multiplier and $\lambda_i c_{ij}$ denotes the total production and transportation cost associated with the order from i^{th} customer served by j^{th} DC. In order to explore the sensitivity of the solution to different levels of response time costs, the multiplier θ is tested for 0.1, 1, 5, 10, 50, 100, and 200 (higher value of θ models the situation in which losing a customer order due to high expected waiting time is extremely costly). In the implementation of the cutting plane method, we start with an a priori set of cuts for the function $f(R) = R/(1+R)$. These cuts are generated based on the piecewise linear approximation $\widehat{f}(R)$ of the function $f(R)$ such that the approximation error (measured by $\widehat{f}(R) - f(R)$) is at most ϵ (see Elhedhli [56]). A similar approach is used by Aboolian et al. [4]. This is in part motivated by our initial computational results which show that the option of starting with an a priori set of cuts improves the performance of the cutting plane method (see Elhedhli [56] for similar results). In all the test problems, we use 32 cuts which corresponds to $\epsilon = 0.001$.

2.8.2 An Illustrative Case Study

Using the first set of test problems (where $I = 50$, $J = 5$, and $K = 3$), we illustrate that the MTO supply chain configuration that considers congestion and its effect on response time can be different from the configuration that ignores congestion. Furthermore, we show empirically that substantial reduction in response times can be achieved with minimal increase in total costs in the design of responsive supply chains.

The first set of test problems consist of 50 customers, 5 potential DC locations ($j = 1$, New York, NY; $j = 2$, Los Angeles, CA; $j = 3$, Chicago, IL; $j = 4$, Houston, TX; $j = 5$, Philadelphia, PA) that can be equipped with 3 capacity levels ($k = 1$, small; $k = 2$, medium; $k = 3$, large). The problem is solved to optimality (with a gap of 10^{-6}) using the cutting plane method. Table 2.1 summarizes the results for different values of the response time cost by setting θ to 0, 0.1, 1, 10, 100, and 1000 under four different cases: M/G/1 case with $CV = 1.5$, M/M/1 case, M/G/1 case with $CV = 0.5$, and M/D/1 case. The table shows the total objective function value (TC), fixed cost (FC), variable cost (VC), total response time cost (RC), total expected waiting time ($\mathbb{E}(W)$), average DC utilization ($\bar{\rho}$), DCs opened and their capacity levels (Open DCs(y_{jk})), expected waiting time at the DCs (W_j), DC utilization (ρ_j), the number of cuts generated (CUT), the number of iterations required (ITR), and the CPU time in seconds (CPU(s)) under different scenarios.

Figure 2.3 shows the effect of changing response time cost on the total workload and capacity of DCs under various scenarios. The supply chain network configuration for two extreme values of response time cost ($\theta = 0$ vs. $\theta = 1000$) are shown in Figure 2.4. Figure 2.5(a) shows the effect of changing response time cost on the total expected waiting time ($\mathbb{E}(W)$), and Figure 2.5(b) shows the effect of changing total expected waiting time $\mathbb{E}(W)$ on the sum of fixed and transportation costs. The observations are as follows:

- Figure 2.4 shows that the supply chain configuration that ignores congestion opens 4 DCs (medium size DCs in New York, Los Angeles, and Chicago and a small DC in Houston) whereas the configuration that considers congestion opens 5 DCs, all with highest capacity level (k=3). Also there is substantial reallocation of customer demand among the DCs in order to balance the workload to reduce the DC utilization and overall response time in the system. Hence, the supply chain configuration that considers congestion and its effect on response time can be different from the traditional configuration that ignores congestion. However, we observe that at very high values of θ, the configurations (location, capacity and demand allocations) are not significantly different among the M/G/1, M/M/1 and M/D/1 cases.

- As we see from Table 2.1, even with very small values of average response time cost, $t = 0.1 \times \text{mean}_{i,j}(\lambda_i c_{ij})$ or $t = 1 \times \text{mean}_{i,j}(\lambda_i c_{ij})$, substantial improvement in the total expected waiting time ($\mathbb{E}[W]$) can be achieved over $t = 0$. For example, for the M/G/1 case ($CV = 1.5$), $\mathbb{E}[W]$ decreases from 77.41 units to 48.37 units for $\theta = 0.1$ and 1 respectively. This is due to the even distribution of demand among DCs. Figure 2.5(a) also shows that the substantial reduction in response time can be achieved with a small values of response time costs. This is because, as we increase the magnitude of the response time cost, DCs with higher capacity are used and/or number of DCs opened increases, average DC utilization decreases, thereby reducing congestion and improving average response time. From Figure 2.5(b), we see that the left portion of the curves are quite flat, indicating that substantial improvement (decrease) in response time can be achieved with a small increase in fixed and transportation costs.

- As the response time cost becomes dominant compared to other cost components, DCs with higher capacity level are opened (See Figure 2.3), average DC utilization decreases, thereby improving (decreasing) the response time. For example, in Table 2.1, in the M/G/1 case ($CV = 1.5$), for $\theta = 1$, 4 medium size DCs are opened, whereas for $\theta = 1000$, 5 large size DCs are opened. The average DC utilization decreases from 0.83 to 0.44 and the expected waiting time decreases from 48.37 to 5.44 units. As θ goes to infinity, the response time costs are *very high* compared to the fixed location cost (FC) and the transportation cost (VC). This results in opening all the potential DCs with the highest capacity levels in an attempt to minimize congestion at DCs. Furthermore, the optimal solution recommends assigning more customers to their closest open facilities (i.e. to a greater extent the closest-assignment property holds). This is evident from Table 2.1(a). However, demand splitting does occur in few cases.

- Table 2.1(a) depicts the allocation of demand to DCs under various scenarios (Closet-assignment of customers to DCs, M/G/1 case with $\theta = 1000$ and M/M/1 case with $\theta = 1000$. For example, in Table 2.1 (a), we observe that the demand for 3 out of 50 customers were split in the M/G/1 case, whereas 5 out of 50 customers were split in the M/M/1 case.

- As we increase the magnitude of the response time cost, the transportation cost decreases initially and then increases. For example, in Table 2.1, in the M/G/1 case ($CV = 1.5$), for $\theta = 0$, VC = 103,577; for $\theta = 1$, VC = 101,494; and for $\theta = 100$, VC = 101,987. This is due to the reallocation of customer demand among DCs in an attempt to reduce the total expected waiting time.

- If transportation costs are *very high* compared to the fixed location cost and the response time costs, the optimal solution recommends assigning more customers to the closest open facilities in most cases. However, demand splitting does occur in few cases.

Table 2.1: Comparison of the MTO supply chain network configurations for M/G/1, M/M/1, and M/D/1 cases: An Illustrative Example.

		M/G/1 (CV = 1.5)	M/M/1 (CV = 0)	M/G/1 (CV = 0.5)	M/D/1
$\theta = 0$	TC	146,965	146,965	146,965	146,965
	FC	43,388	43,388	43,388	43,388
	VC	103,577	103,577	103,577	103,577
	RC	0	0	0	0
	$\mathbb{E}(W)$	∞	∞	∞	∞
	$\bar{\rho}$	0.96	0.96	0.96	0.96
	Open DCs	1(2), 2(2), 3(2), 4(1)	1(2), 2(2), 3(2), 4(1)	1(2), 2(2), 3(2), 4(1)	1(2), 2(2), 3(2), 4(1)
	W_j	[138.79, ∞, 5.27, ∞]	[138.79, ∞, 5.27, ∞]	[138.79, ∞, 5.27, ∞]	[138.79, ∞, 5.27, ∞]
	ρ_j	[0.99, 1.00, 0.84, 1.00]	[0.99, 1.00, 0.84, 1.00]	[0.99, 1.00, 0.84, 1.00]	[0.99, 1.00, 0.84, 1.00]
	CUT	0	0	0	0
	ITR	1	1	1	1
	CPU(s)	0.14	0.14	0.14	0.14
$\theta = 0.1$	TC	148,567	148,333	148,042	146,724
	FC	46,816	43,388	43,388	43,388
	VC	101,494	104,235	104,093	104,033
	RC	257	710	560	503
	$\mathbb{E}(W)$	77.41	214.05	169.00	152.40
	$\bar{\rho}$	0.83	0.96	0.96	0.96
	Open DCs	1(2), 2(2), 3(2), 4(2)	1(2), 2(2), 3(2), 4(1)	1(2), 2(2), 3(2), 4(1)	1(2), 2(2), 3(2), 4(1)
	W_j	[10.09, 33.51, 2.49, 2.83]	[33.93, 98.34, 7.24, 74.55]	[42.56, 124.89, 6.66, 94.06]	[48.12, 139.37, 6.43, 107.05]
	ρ_j	[0.99, 0.97, 0.71, 0.74]	[0.97, 0.99, 0.88, 0.99]	[0.98, 0.99, 0.87, 0.99]	[0.98, 0.99, 0.87, 0.99]
	CUT	4	20	20	20
	ITR	2	6	6	6
	CPU(s)	0.35	0.88	1.69	0.9
$\theta = 1$	TC	150,269	149,683	149,286	149139
	FC	46,816	46,816	46,816	46816
	VC	101,849	101,744	101,649	101,609
	RC	1,604	1,124	822	714
	$\mathbb{E}(W)$	48.37	33.9	24.79	21.55
	$\bar{\rho}$	0.83	0.83	0.83	0.83
	Open DCs	1(2), 2(2), 3(2), 4(2)	1(2), 2(2), 3(2), 4(2)	1(2), 2(2), 3(2), 4(2)	1(2), 2(2), 3(2), 4(2)
	W_j	[10.09, 15.08, 2.49, 3.39]	[10.09, 18.11, 2.49, 3.21]	[10.09, 22.02, 2.49, 3.06]	[10.09, 24.19, 2.49, 3.00]
	ρ_j	[0.91, 0.94, 0.71, 0.77]	[0.91, 0.95, 0.71, 0.76]	[0.91, 0.96, 0.71, 0.75]	[0.91, 0.96, 0.71, 0.75]
	CUT	4	8	16	12
	ITR	2	3	5	4
	CPU(s)	0.33	0.41	0.91	0.45
$\theta = 10$	TC	157,274	155,674	154,324	153874
	FC	52,078	49,447	49,447	49447
	VC	101,407	101,640	101,638	101616
	RC	3,789	4,587	3,240	2811
	$\mathbb{E}(W)$	11.43	13.84	9.77	8.48
	$\bar{\rho}$	0.67	0.75	0.75	0.75
	Open DCs	1(3), 2(3), 3(2), 4(2)	1(2), 2(3), 3(2), 4(2)	1(2), 2(3), 3(2), 4(2)	1(2), 2(3), 3(2), 4(2)
	W_j	[1.75, 2.27, 2.10, 1.94]	[6.63, 2.27, 2.54, 2.39]	[6.63, 2.27, 2.58, 2.36]	[6.78, 2.27, 2.65, 2.28]
	ρ_j	[0.64, 0.69, 0.68, 0.66]	[0.87, 0.69, 0.72, 0.71]	[0.87, 0.69, 0.72, 0.70]	[0.87, 0.69, 0.73, 0.69]
	CUT	12	12	8	12
	ITR	4	4	3	4
	CPU(s)	0.65	0.62	0.49	0.60
$\theta = 100$	TC	182,451	176,255	172,486	170,940
	FC	57,340	57,340	57,340	54,709
	VC	101,987	101,494	101,494	101,557
	RC	23,124	17,421	13,650	14,675
	$\mathbb{E}(W)$	6.98	5.26	4.12	4.43
	$\bar{\rho}$	0.56	0.56	0.56	0.61
	Open DCs	1(3), 2(3), 3(3), 4(3)	1(3), 2(3), 3(3), 4(3)	1(3), 2(3), 3(3), 4(3)	1(3), 2(3), 3(3), 4(2)
	W_j	[1.45, 1.68, 0.96, 1.05]	[1.54, 1.84, 0.91, 0.97]	[1.54, 1.84, 0.91, 0.97]	[1.75, 2.15, 1.00, 1.54]
	ρ_j	[0.59, 0.63, 0.49, 0.51]	[0.61, 0.65, 0.48, 0.49]	[0.61, 0.65, 0.48, 0.49]	[0.64, 0.68, 0.50, 0.61]
	CUT	16	4	4	16
	ITR	5	2	2	5
	CPU(s)	0.77	0.26	0.35	0.93
$\theta = 1000$	TC	358,156	316,316	289,863	280,873
	FC	71,675	71,675	71,675	71,675
	VC	106,029	102,974	99,683	99,460
	RC	180,453	141,666	118,506	109,738
	$\mathbb{E}(W)$	5.44	4.27	3.57	3.31
	$\bar{\rho}$	0.44	0.44	0.44	0.44
	Open DCs	1(3), 2(3), 3(3), 4(3), 5(3)	1(3), 2(3), 3(3), 4(3), 5(3)	1(3), 2(3), 3(3), 4(3), 5(3)	1(3), 2(3), 3(3), 4(3), 5(3)
	W_j	[0.64, 1.39, 0.78, 0.84, 0.55]	[0.66, 1.51, 0.77, 0.83, 0.50]	[0.68, 1.67, 0.76, 0.84, 0.43]	[0.72, 1.67, 0.76, 0.85, 0.41]
	ρ_j	[0.39, 0.58, 0.44, 0.46, 0.35]	[0.40, 0.60, 0.44, 0.45, 0.34]	[0.40, 0.63, 0.43, 0.46, 0.30]	[0.42, 0.63, 0.43, 0.46, 0.29]
	CUT	35	35	25	10
	ITR	8	8	6	3
	CPU(s)	1.66	1.45	1.2	0.39

θ: Multiplier for the response time costs;
TC: Total cost; FC: Fixed cost; VC: Variable cost; RC: Response time cost;
$\mathbb{E}(W)$: Total expected waiting time;
$\bar{\rho}$: Average DC utilization;
Open DCs (y_{kl}): DCs selected open and their capacity levels;
W_j: Response time;
ρ_j: DC utilization;
CUT: Number of cuts generated;
ITR: Number of iterations;
CPU(s): CPU time in sec.

Table 2.1(a): Allocation of customer demand to DCs under various scenarios

	DC Opened	Closest-Assignment of Customers to DCs without Demand Splitting	Demand Allocated, M/G/1 Case (CV = 1.5), Θ = 1000	Demand Allocated, M/M/1 Case, Θ = 1000
1	New York NY (Large)	New York NY, Boston MA, VA Beach VA	*New York NY (0.99)*	New York NY(1)
2	Los Angeles CA (Large)	Los Angeles CA, Phoenix AZ, San Diego CA, San Jose CA, San Francisco CA, Denver CO, Seattle WA, Portland OR, Las Vegas NV, Tucson AZ, Long Beach CA, Albuquerque NM, Fresno CA, Mesa AZ, Oakland CA, Santa Ana CA	Los Angeles CA, Phoenix AZ, San Diego CA, San Jose CA, San Francisco CA, Seattle WA, Portland OR, Las Vegas NV, Long Beach CA, Fresno CA, *Mesa AZ (0.10),* Oakland CA, Santa Ana CA	Los Angeles CA, Phoenix AZ, San Diego CA, San Jose CA, San Francisco CA, Seattle WA, Portland OR, Las Vegas NV, Boston MA(0.06), Tucson AZ (0.08), Long Beach CA, Fresno CA, *Mesa AZ (1),* Oakland CA, Santa Ana CA
3	Chicago IL (Large)	Chicago IL, Detroit MI, Indianapolis IN, Columbus OH, Memphis TN, Milwaukee WI, Nashville TN, Cleveland OH, KS City MO, Atlanta GA, Omaha NE, Minneapolis MN, Saint Louis MO	Chicago IL, Detroit MI (0.96), Indianapolis IN, Memphis TN, Milwaukee WI, Nashville TN, Denver CO, KS City MO, Omaha NE, Minneapolis MN, Saint Louis MO, Wichita KS	Chicago IL, Detroit MI (1), Indianapolis IN, Memphis TN, Milwaukee WI, Columbus OH(0.28), Nashville TN, Denver CO (0.44), KS City MO, Omaha NE, Minneapolis MN, Saint Louis MO, Wichita KS
4	Houston TX (Large)	Houston TX, Dallas TX, San Antonio TX, Austin TX, El Paso TX, Fort Worth TX, Tucson AZ, Oklahoma City OK , New Orleans LA, Tulsa OK, Colorado Springs CO, Miami FL, Wichita KS	Houston TX, Dallas TX, San Antonio TX, Austin TX, El Paso TX, Fort Worth TX, Tucson AZ, Oklahoma City OK , New Orleans LA, Albuquerque NM, *Mesa AZ(0.90)*, Tulsa OK, Colorado Springs CO	Houston TX, Dallas TX, San Antonio TX, Austin TX, El Paso TX, Fort Worth TX, Tucson AZ(0.92), Oklahoma City OK , Denver CO(0.54), New Orleans LA, Albuquerque NM, Tulsa OK, Colorado Springs CO
5	Philadelphia PA (Large)	Philadelphia PA, Jacksonville FL, Baltimore MD, Washington DC, Charlotte NC	*New York NY (0.01),* Philadelphia PA, *Detroit MI (0.04)*, Jacksonville FL, Columbus OH, Baltimore MD, Boston MA, Washington DC, Charlotte NC, Cleveland OH, VA Beach VA, Atlanta GA, Miami FL	Philadelphia PA, Detroit MI (0), Jacksonville FL, Columbus OH(0.72), Baltimore MD, Boston MA(0.94), Washington DC, Charlotte NC, Cleveland OH, VA Beach VA, Atlanta GA, Miami FL

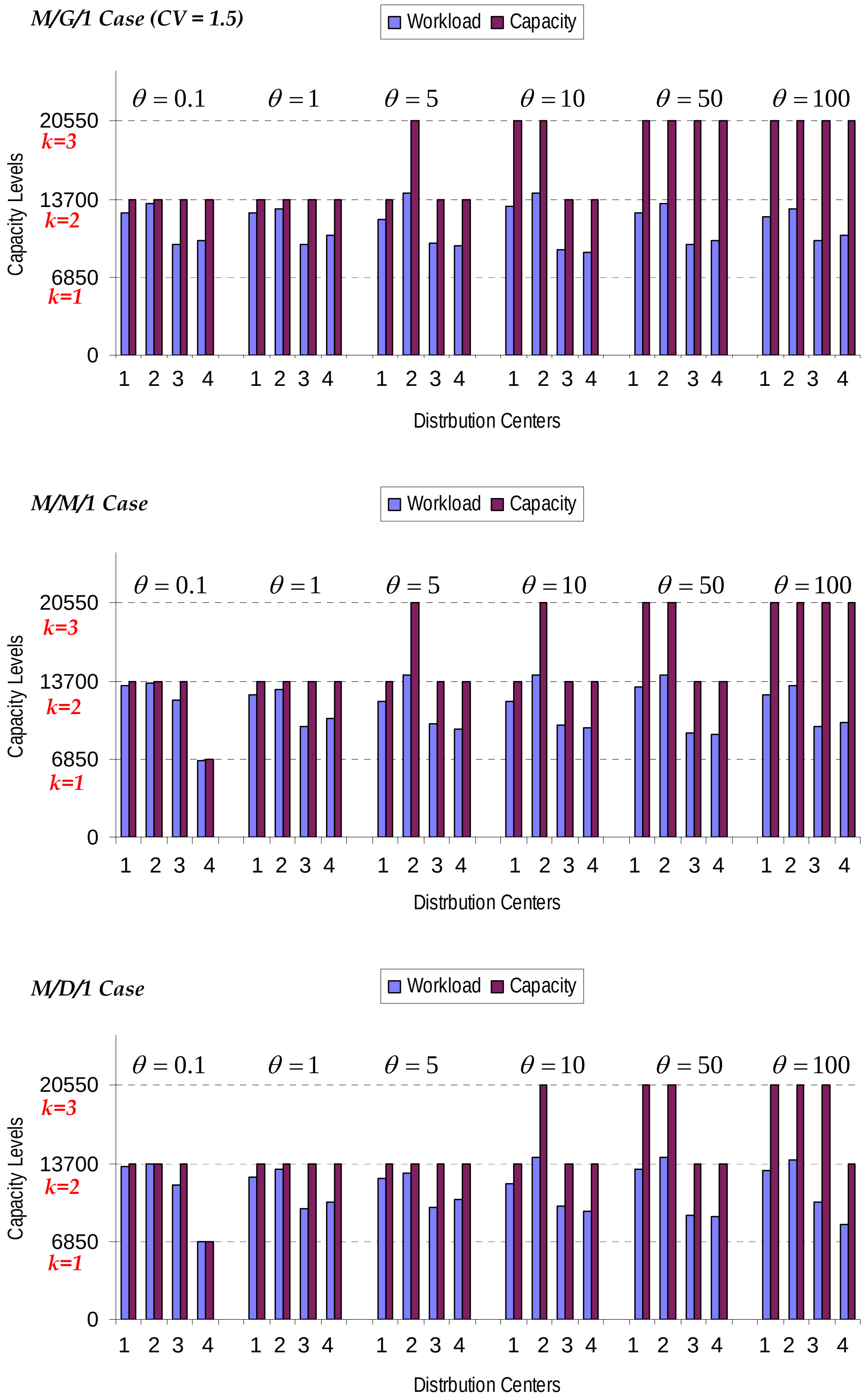
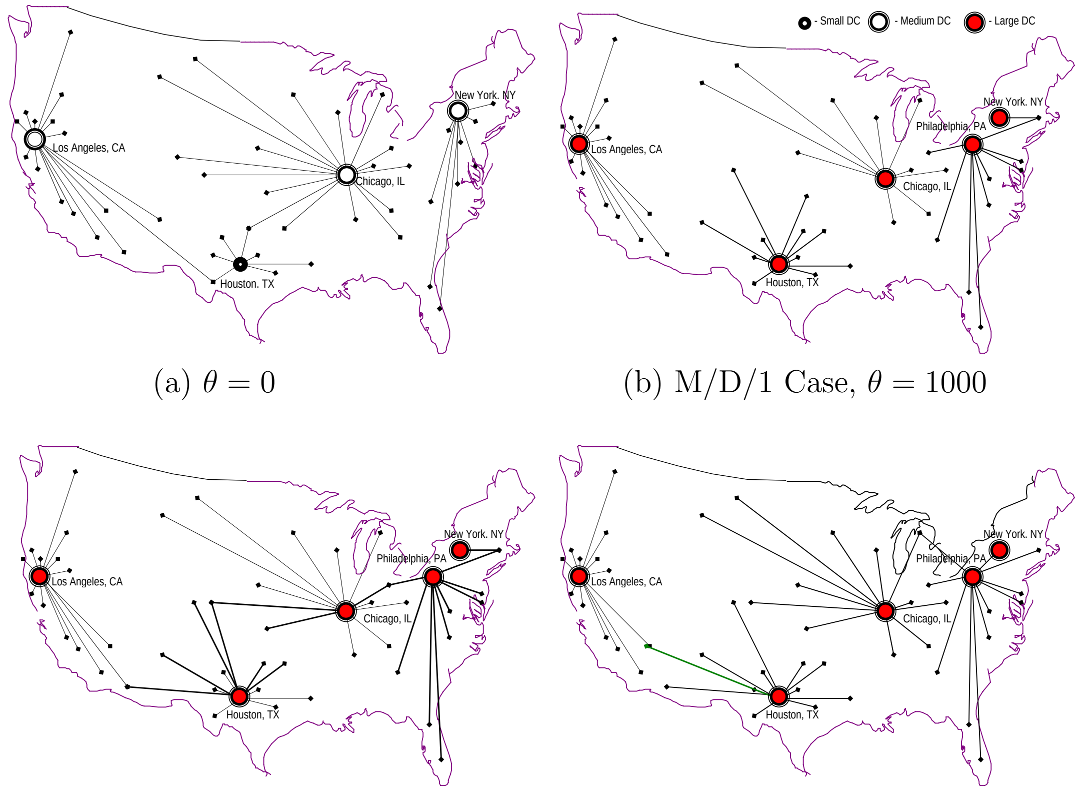


Figure 2.3: Effect of changing response time cost on the total workload and capacity of DCs under various scenarios



(a) $\theta = 0$ (b) M/D/1 Case, $\theta = 1000$

(c) M/M/1 Case, $\theta = 1000$ (d) M/G/1 Case $(CV = 1.5)$, $\theta = 1000$

Figure 2.4: Effect of changing response time cost on the supply chain network configuration

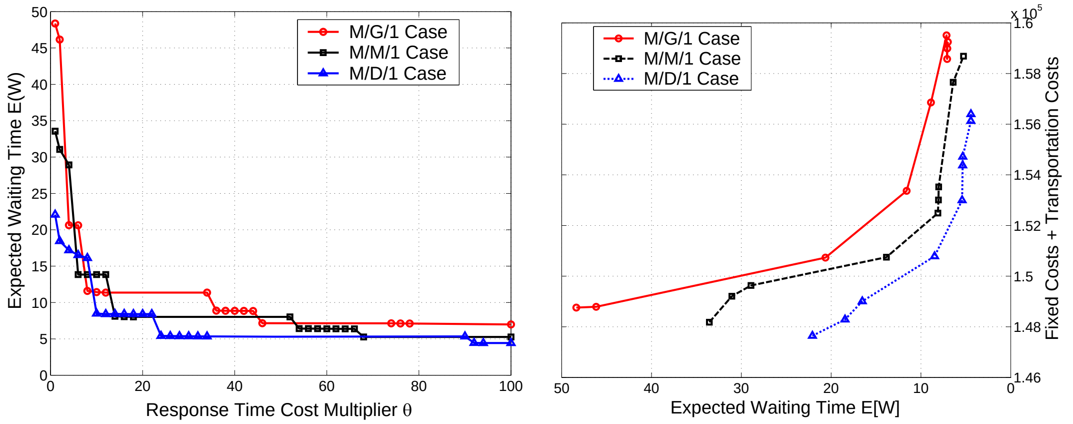


Figure 2.5: (a) Effect of changing response time cost (t) on the total expected waiting time $\mathbb{E}[W]$, (b) Effect of changing expected waiting time $\mathbb{E}[W]$ on the fixed costs + transportation costs

2.8.3 Performance of the Cutting Plane Method

Table 2.2 displays the performance of the cutting plane method for MTO supply chain design problems under M/G/1 ($CV = 1.5$), M/M/1 ($CV = 1.0$), and M/D/1 ($CV = 0$) cases for varying problem sizes. The columns marked FC, VC, and RC represent the fixed costs, the variable production and transportation, and the response time costs respectively, expressed as a percentage of total costs (TC). The columns marked $\mathbb{E}(W)$ is the total average waiting time in the system, $\overline{\rho}$ is the average DC utilization, and DC represents the number of DCs opened. The table also displays the number of constraints generated (CUT), the number of iterations of the method (ITR), and the total CPU time in seconds required to obtain the optimal solution.

The results show that the average CPU time for M/G/1, M/M/1 and M/D/1 cases are 48, 21, and 9 seconds respectively, whereas the average number of cuts required are 26, 25, and 25 respectively. Also the maximum CPU time for M/G/1, M/M/1 and M/D/1 cases are 1147, 410, and 107 seconds respectively, whereas the maximum number of cuts required are 66, 60, and 50 respectively. The computation times reveal the stability and the efficiency of the cutting plane method for different percentages of fixed, variable and response time costs, whereas the number of iterations imply that only a fraction of the constraints in $[P_L(H)]$ are required. As the magnitude of response time cost (t) increases, the percentage of response time cost becomes more significant with respect to other cost components and the method seems to require more CPU time and iterations as large number of cuts are generated. It is also worthwhile noting that the computational times for the second set of instances (I=50, J = 20, and K=3) are comparatively higher than others because the optimal solution has highly congested DCs. In the model, this corresponds to the value of $R/(1 + R)$ approaching 1. At the flat portion of $R/(1 + R)$, a higher number of cuts is needed to close the gap. Furthermore, in almost all of the instances, the M/G/1 case requires more cuts, and hence more CPU time to solve than the M/M/1 and M/D/1 cases. This is attributed to the nonlinearity in the expression of expected waiting time for M/G/1 queues.

Table 2.2: Computational performance of the cutting plane method: MTO supply chain design

					M/G/1 Case (CV = 1.5)										M/M/1 Case										M/D/1 Case									
No.	I	J	K	θ	FC (%)	VC (%)	RC (%)	TC	E(W)	$\overline{\rho}$	DC	CUT	ITR	CPU (s)	FC (%)	VC (%)	RC (%)	TC	E(W)	$\overline{\rho}$	DC	CUT	ITR	CPU (s)	FC (%)	VC (%)	RC (%)	TC	E(W)	$\overline{\rho}$	DC	CUT	ITR	CPU (s)
1	50	10	3	0.1	40	60	0	133,713	221	0.93	6	24	5	2	40	60	0	133,594	152	0.93	6	18	4	1	40	60	0	133,484	95	0.92	6	18	4	1
				1	39	59	2	136,202	198	0.93	6	30	6	2	39	59	1	135,181	126	0.93	6	24	5	1	40	60	1	134,356	66	0.93	6	24	5	1
				5	39	58	3	140,635	60	0.85	6	24	5	2	40	58	2	139,148	41	0.85	6	18	4	2	40	59	1	137,862	26	0.86	6	30	6	3
				10	40	56	3	143,888	35	0.77	6	24	5	3	39	57	4	141,829	37	0.85	6	30	6	3	40	58	2	139,498	23	0.85	6	18	4	2
				50	42	52	6	158,723	14	0.61	6	6	2	2	40	53	8	153,355	17	0.70	6	18	4	3	39	54	6	148,628	14	0.77	6	18	4	2
				100	39	49	12	168,467	14	0.61	6	18	4	5	40	51	9	162,917	10	0.61	6	6	2	2	39	52	9	156,215	11	0.7	6	18	4	3
2	50	20	3	0.1	59	41	0	122,775	419	0.92	11	44	5	10	59	41	0	122,625	296	0.92	11	44	5	8	59	41	0	122,483	188	0.92	11	44	5	7
				1	56	43	1	124,843	198	0.91	10	40	5	10	56	43	1	124,153	139	0.91	10	40	5	7	59	41	1	123,482	99	0.92	11	44	5	7
				5	57	39	4	129,842	117	0.86	11	55	6	24	58	39	3	127,937	80	0.87	11	55	6	12	55	42	2	126,045	63	0.91	10	40	5	8
				10	57	37	6	134,288	89	0.83	11	55	6	44	57	38	5	131,372	74	0.86	11	55	6	22	58	39	3	128,366	45	0.87	11	44	5	10
				50	51	37	12	154,123	41	0.71	10	50	6	541	53	37	10	147,618	33	0.74	10	50	6	281	54	38	8	140,915	24	0.76	10	40	5	50
				100	52	33	16	170,689	29	0.64	11	66	7	1147	49	35	15	160,786	27	0.71	10	60	7	410	52	37	12	150,932	20	0.74	10	50	6	107
3	100	5	3	1	27	73	0	196,048	51	0.83	4	12	4	1	27	73	0	195,869	37	0.83	4	8	3	1	27	73	0	195,713	21	0.83	4	12	4	1
				5	27	72	1	197,594	41	0.83	4	4	2	0	27	73	1	196,932	26	0.83	4	4	2	0	27	73	0	196,389	16	0.83	4	12	4	1
				10	26	72	2	199,377	37	0.83	4	16	5	1	27	72	1	198,142	26	0.83	4	12	4	1	27	73	1	197,085	15	0.83	4	4	2	0
				50	28	69	3	206,964	12	0.67	4	16	5	1	27	70	3	204,858	13	0.75	4	12	4	1	26	71	3	202,211	13	0.83	4	8	3	1
				100	28	67	5	212,399	12	0.67	4	16	5	1	28	68	4	209,168	8	0.67	4	12	4	1	27	69	4	206,235	8	0.75	4	12	4	1
				200	29	65	6	220,249	7	0.56	4	16	5	1	28	66	5	216,430	6	0.61	4	8	3	1	28	68	5	211,637	6	0.67	4	12	4	1
4	100	10	3	1	47	52	0	165,402	134	0.74	8	32	5	8	34	66	1	178,665	284	0.93	6	30	6	5	34	66	0	178,135	153	0.93	6	24	5	3
				5	53	47	0	166,564	32	0.65	9	18	3	6	37	62	1	180,104	57	0.86	7	21	4	3	38	62	0	179,576	38	0.86	7	21	4	3
				10	52	47	1	167,152	31	0.65	9	27	4	6	37	62	1	181,077	49	0.85	7	28	5	3	37	62	1	180,216	32	0.86	7	21	4	2
				50	51	46	3	171,453	28	0.65	9	36	5	9	38	60	3	186,480	26	0.77	7	28	5	4	37	61	2	184,155	25	0.85	7	28	5	3
				100	50	44	6	176,749	28	0.65	9	27	4	12	37	58	5	191,084	24	0.77	7	28	5	5	38	59	3	187,494	15	0.77	7	28	5	4
				200	48	46	6	184,572	14	0.55	8	8	2	12	35	59	6	198,182	16	0.70	6	18	4	4	34	61	5	192,882	13	0.77	6	24	5	4
5	100	20	3	1	47	52	0	165,218	159	0.74	8	32	5	9	47	52	0	165,061	119	0.74	8	32	5	6	48	52	0	164,903	79	0.74	8	24	4	5
				5	47	52	1	166,275	86	0.74	8	32	5	9	47	52	0	165,867	65	0.74	8	24	4	5	47	52	0	165,458	44	0.74	8	32	5	6
				10	53	47	0	166,751	32	0.65	9	18	3	5	53	47	0	166,503	23	0.65	9	18	3	4	47	52	1	165,939	36	0.74	8	24	4	5
				50	52	46	2	169,629	29	0.65	9	36	5	7	52	47	1	168,556	20	0.65	9	27	4	5	52	47	1	167,667	13	0.65	9	27	4	5
				100	51	45	4	173,133	28	0.65	9	36	5	10	51	46	3	170,995	20	0.65	9	27	4	5	52	46	2	169,281	13	0.65	9	27	4	5
				200	46	47	7	179,780	25	0.65	8	32	5	21	50	45	6	175,861	20	0.65	9	36	5	9	51	46	4	172,442	13	0.65	9	27	4	5
6	150	5	5	1	24	76	0	225,407	149	0.91	4	16	5	2	24	76	0	225,124	114	0.91	4	16	5	2	24	76	0	224,830	81	0.92	4	16	5	2
				5	24	75	1	227,400	96	0.91	4	8	3	1	24	76	1	226,563	64	0.91	4	12	4	1	24	76	0	225,807	42	0.91	4	16	5	2
				10	25	74	1	229,437	38	0.82	4	8	3	1	24	75	1	227,963	59	0.91	4	16	5	2	24	76	1	226,646	34	0.91	4	12	4	1
				50	25	73	2	234,236	19	0.75	4	16	5	2	26	73	1	232,794	13	0.75	4	4	2	1	25	74	1	230,868	13	0.82	4	16	5	2
				100	25	72	3	238,462	18	0.75	4	12	4	1	25	72	2	235,758	13	0.75	4	16	5	2	26	73	2	233,477	8	0.75	4	4	2	1
				200	26	70	4	243,902	12	0.67	4	12	4	2	26	71	3	240,689	8	0.67	4	8	3	1	25	72	3	237,123	8	0.75	4	16	5	2
7	150	10	5	1	31	69	0	207,164	294	0.92	6	24	5	6	31	69	0	206,933	224	0.93	6	24	5	5	31	69	0	206,707	144	0.93	6	24	5	6
				5	32	68	0	208,307	87	0.88	6	6	2	3	34	66	0	207,975	88	0.89	7	28	5	5	31	69	0	207,474	80	0.92	6	24	5	5
				10	32	68	1	209,108	81	0.88	6	18	4	5	32	68	1	208,534	55	0.88	6	6	2	2	34	66	0	208,034	47	0.89	7	28	5	6
				50	34	64	2	213,571	54	0.83	7	28	5	10	34	64	2	211,759	37	0.84	7	28	5	6	35	64	1	210,229	23	0.84	7	21	4	4
				100	31	65	3	217,917	37	0.8	6	24	5	13	34	63	3	215,170	35	0.83	7	21	4	7	34	64	2	212,314	22	0.84	7	21	4	5
				200	32	64	4	223,722	25	0.73	6	18	4	21	31	65	4	220,220	24	0.8	6	24	5	16	34	63	4	216,271	21	0.83	7	21	4	7
8	150	20	5	1	48	52	0	177,444	332	0.9	10	40	5	11	48	52	0	177,277	245	0.9	10	40	5	8	48	52	0	177,113	162	0.9	10	40	5	8
				5	45	54	1	178,543	169	0.89	9	27	4	11	45	54	0	178,141	115	0.89	9	27	4	7	48	52	0	177,693	99	0.9	10	40	5	9
				10	45	54	1	179,538	157	0.89	9	36	5	16	45	54	1	178,818	103	0.89	9	27	4	8	45	54	0	178,190	61	0.89	9	27	4	7
				50	46	52	2	184,593	62	0.81	9	27	4	38	45	53	2	183,147	64	0.85	9	36	5	27	45	53	2	180,893	51	0.88	9	36	5	14
				100	45	51	4	188,335	59	0.81	9	27	4	54	45	52	3	185,827	41	0.81	9	27	4	27	45	52	2	183,620	36	0.85	9	36	5	24
				200	49	47	5	194,476	37	0.71	10	30	4	211	44	51	5	190,699	39	0.81	9	27	4	68	45	52	3	186,730	24	0.81	9	27	4	25
				min	24	33	0	122,775	7	0.55	4	4	2	0	24	35	0	122,625	6	0.61	4	4	2	0	24	37	0	122,483	6	0.65	4	4	2	0
				max	59	76	16	243,902	419	0.93	11	66	7	1147	59	76	15	240,689	296	0.93	11	60	7	410	59	76	12	237,123	188	0.93	11	50	6	107
				mean	41	56	3	183,022	82	0.77	7	26	4	48	39	58	3	182,995	67	0.80	7	25	4	21	39	59	2	186,516	46	0.82	7	25	4	9

I: No. of customers; J: No. of potential DCs; K: No. of capacity levels at each DC;
θ: Multiplier for the response time cost;
FC: Fixed cost; VC: Variable cost; RC: Response time cost; TC: Total cost (Note that FC, VC and RC are expressed as percentages of the total cost, TC);
$\mathbb{E}(W)$: Total Expected Waiting Time; $\bar{\rho}$: Average DC utilization; DC: No. of DCs selected open;
CUT: Number of cuts generated; ITR: Number of iterations; CPU(s): CPU time in sec.

2.8.4 Performance of the Lagrangean Heuristic

We test the Lagrangean heuristic on the same set of test problems used in the previous section and report the performance in Table 2.3. The Lagrangean bound (LAG) is expressed as a percentage of the optimal solution. The quality of the heuristic solution (GAP) is expressed as a percentage of the optimal solution: 100 × (Heuristic Solution − Optimal Solution)/Optimal Solution. In all these test problems, the heuristic is activated at the final iteration of the Lagrangean procedure. The table also reports the computational time of the subproblems (SP), the master problem (MP) and the heuristic (H) as a percentage of the total computational time (CPU) for various instances. The heuristic succeeds in finding feasible solutions that are within a maximum of 4.90%, 4.67%, and 4.99% from the optimal solution for the M/G/1, M/M/1, and M/D/1 case respectively. On average the gap is 3.17%, 2.45%, and 2.73% from the optimal solution for the M/G/1, M/M/1, and M/D/1 cases respectively. In terms of computational time, the proposed heuristic takes an average of 195, 175, and 150 sec for M/G/1, M/M/1, and M/D/1 cases respectively for problems with up to 150 customers, 20 DCs and 5 capacity levels. The total computational time can be as high as 410 sec in some cases. Also the computational times seem to depend on the coefficient of variation of service times. In most of the cases, the M/G/1 instances (CV = 1.5) takes more time than M/M/1, and M/D/1 cases. It is evident that the solution of the subproblems accounts for most of the computational time: 81.83%, 83.88%, and 86.18% for M/G/1 case, M/M/1, and M/D/1 cases respectively. The master problem accounts for 17.26%, 15.08%, 12.85%, for M/G/1 case, M/M/1, and M/D/1 cases respectively whereas the heuristic accounts for 0.91%, 1.05%, 0.97% on average. This supports the claim that the difficulty of the original problem has been transferred to the subproblems, while keeping them still computationally tractable. This pays off in getting a lower bound and a heuristic solution close to optimal solution.

In Table 2.4, we compare the performance of the cutting plane method and the Lagrangean heuristic for high response time costs (by setting θ to 400, 600, 800, 1000, 1500, 2000). The results show that as the Lagreangean heursitic is computationally efficient compared to the cutting plane method for intances where the response time cost component is dominant.

2.9 Extensions

We present extensions to our original model to consider systems with multiple customer classes and general distributions of demand and service times. We show that our solution procedures can be easily extended to deal with these cases. Later, in Chapter 4, we consider a system with multiple customer classes where we incorporate a probability constraint that ensures that waiting time does not exceed a pre-specified threshold.

2.9.1 Systems with Multiple Customer Classes

In certain settings, the demand for finished product may arise from a different customer classes. Assume that these customers belong to one of N *priority classes* (class 1 has the highest priority and class N has the lowest) and whenever a DC becomes free to serve a new customer from the queue, the customer selected for service is the member of the highest priority class who has waited longest in the queue. In other words, customers are selected to begin service in the order of priority classes, but on a first-come-first-serve (FCFS) basis within each priority class. Let the system incur a delay cost t_n per unit time one of its customers spends in the system (either waiting in the queue or in service). Without the loss of generality, we assume that the average response time costs per unit time are ordered as follows: $t_1 \geq t_2 \geq t_3 \geq ... \geq t_n$. We assume that the expected service time $1/\mu_{jk}$ is same for all priority classes that are assigned to a DC j equipped with capacity level k. If x_{ij}^n denote the fraction of demand for the product from customer i of class n served by DC j $(0 \leq x_{ij}^n \leq 1)$, then x_{ij}^ns must satisfy the constraint

$$\sum_{i=1}^{I}\sum_{n=1}^{N}\lambda_i^n x_{ij}^n \leq \sum_{k=1}^{K}\mu_{jk}y_{jk} \qquad \forall j \tag{2.21}$$

$$\sum_{j=1}^{J} x_{ij}^n = 1 \qquad \forall i, n \tag{2.22}$$

Table 2.3: Computational performance of the Lagrangean heuristic: MTO supply chain design

					M/G/1 Case (CV = 1.5)						M/M/1 Case						M/D/1 Case					
No.	I	J	K	θ	LAG(%)	GAP(%)	SP(%)	MP(%)	H(%)	CPU(s)	LAG(%)	GAP(%)	SP(%)	MP(%)	H(%)	CPU(s)	LAG(%)	GAP(%)	SP(%)	MP(%)	H(%)	CPU(s)
1	50	10	3	0.1	95.73	1.30	83.08	16.30	0.62	19	92.42	0.62	84.65	13.74	1.61	15	94.52	0.3	86.54	12.52	0.94	12
				1	99.65	1.65	88.55	10.86	0.59	33	98.73	2.72	80.38	18.1	1.52	21	98.80	1.53	91.51	8.25	0.24	24
				5	98.82	4.51	83.32	14.90	1.78	37	94.38	3.82	89.02	10.3	0.68	31	99.53	3.31	84.62	15.11	0.27	28
				10	92.40	1.82	78.1	21.34	0.56	38	99.15	3.25	79.38	19.45	1.17	39	98.66	3.85	84.76	14.43	0.81	31
				50	99.59	0.68	84.96	13.54	1.50	39	94.70	1.00	86.13	12.39	1.48	65	98.56	0.7	87.25	10.92	1.83	34
				100	90.10	3.77	78.99	20.87	0.14	49	98.82	1.63	82.43	16.8	0.77	70	98.21	1.19	81.48	16.64	1.88	34
2	50	20	3	0.1	93.29	1.91	76.15	22.89	0.96	15	99.80	2.65	87.27	12.61	0.12	12	97.59	2.17	83.74	14.64	1.62	12
				1	98.08	3.26	80.91	17.57	1.52	32	94.52	1.54	78.89	19.14	1.97	30	95.52	3.97	81.84	17.33	0.83	16
				5	93.82	1.87	75.72	22.99	1.29	47	98.92	2.48	79.65	19.82	0.53	37	95.94	4.56	87.87	11.02	1.11	34
				10	96.48	1.15	75.29	23.70	1.01	48	97.30	3	87.85	11.95	0.2	51	97.62	2.16	89.49	9.38	1.13	41
				50	92.88	2.97	77.14	22.68	0.18	49	94.20	1.52	79.26	19.8	0.94	64	97.91	1.35	89.09	9.45	1.46	41
				100	95.49	1.93	87.16	12.55	0.29	53	96.65	3.08	78.44	19.79	1.77	78	98.28	1.14	84	15.12	0.88	41
3	100	5	3	1	95.95	4.69	81.62	17.79	0.59	115	92.73	1.17	85.13	13.93	0.94	100	98.19	1.61	90.91	8.63	0.46	103
				5	99.47	3.87	78.5	19.67	1.83	118	95.25	3.52	83.67	16.1	0.23	102	97.53	3.75	86.89	13.04	0.07	109
				10	95.29	4.48	80.74	18.35	0.91	119	95.12	0.51	82.1	16.63	1.27	112	96.25	4.82	87.61	11.15	1.24	111
				50	99.73	3.36	78.03	20.18	1.79	120	98.21	0.24	85.81	13.97	0.22	125	98.15	2.18	84.81	15.04	0.15	113
				100	90.80	3.92	76.08	23.45	0.47	120	94.52	3.62	81.01	17.01	1.98	133	94.62	3.06	88.65	10.36	0.99	116
				200	92.57	2.59	82.48	16.25	1.27	123	96.54	3.13	80.39	19.35	0.26	152	98.11	3.47	89.47	10.02	0.51	118
4	100	10	3	1	91.10	4.41	82.27	15.99	1.74	214	92.87	2.75	82.56	17.29	0.15	150	95.74	3.13	89.72	9.28	1.00	167
				5	96.89	3.55	87.18	12.33	0.49	216	98.66	4.5	85.73	14.13	0.14	160	96.23	3.46	82.13	16.49	1.38	181
				10	93.28	0.84	86.65	12.16	1.19	219	93.25	0.44	79.59	18.66	1.75	166	96.32	3.06	90.64	8.95	0.41	187
				50	98.34	2.62	79.28	18.73	1.99	221	95.27	2.57	84.8	13.19	2.01	193	99.80	3.19	89.14	9.48	1.38	187
				100	97.79	4.27	77.43	21.00	1.57	221	94.60	2.71	85.59	13.63	0.78	196	94.91	3.84	81.19	17.57	1.24	188
				200	98.45	3.32	87.67	11.66	0.67	224	96.22	0.66	85.97	13.05	0.98	205	94.55	3.08	84.42	14.60	0.98	192
5	100	20	3	1	95.46	4.43	83.19	15.41	1.40	204	95.60	1.84	86.92	11.79	1.29	407	94.15	2.65	89.25	9.39	1.36	334
				5	95.14	0.75	79.38	19.91	0.71	217	96.18	3.42	80.78	18.64	0.58	125	97.34	4.38	89.88	8.30	1.82	193
				10	97.15	0.39	82.72	16.27	1.01	220	98.96	2.45	88.25	10.55	1.2	280	98.10	0.86	85.36	13.21	1.43	123
				50	98.95	4.26	87.52	12.35	0.13	315	99.90	1.49	78.55	19.46	1.99	322	94.38	0.84	88	11.74	0.26	221
				100	98.59	4.77	86.89	11.55	1.56	245	94.38	2.26	85.74	14	0.26	333	95.31	3.05	85.97	12.69	1.34	242
				200	91.13	4.37	84.83	15.03	0.14	177	98.34	2.38	87.91	11.15	0.94	365	94.02	4.18	87.31	12.02	0.67	193
6	150	5	5	1	93.31	4.59	82.87	15.70	1.43	410	95.44	4.47	86.34	12.03	1.63	291	95.25	0.92	86.41	12.58	1.01	255
				5	92.45	3.78	80.2	19.47	0.33	333	94.98	3.13	85.63	13.85	0.52	189	97.55	4.64	85.81	13.92	0.27	348
				10	96.65	3.67	86.74	11.85	1.41	393	98.57	1.73	80.54	18.36	1.1	232	94.95	0.39	85.28	13.63	1.09	129
				50	91.01	4.03	74.3	24.99	0.71	206	96.79	4.38	82.89	16.48	0.63	304	98.73	4.19	83.66	14.77	1.57	156
				100	91.30	4.21	82.07	17.31	0.62	245	97.69	3.73	80.29	19.17	0.54	144	97.74	4.99	81.59	17.56	0.85	154
				200	93.77	1.28	89.16	10.25	0.59	378	99.33	0.61	83.17	15.74	1.09	385	95.88	1.15	85.73	14.15	0.12	163
7	150	10	5	1	93.40	3.74	80.33	18.47	1.20	355	93.00	4.67	81.72	17.1	1.18	273	99.96	3.77	86.23	11.78	1.99	324
				5	95.63	4.87	75.13	24.62	0.25	154	95.24	2.15	85.39	13.78	0.83	132	97.98	1.64	89.47	9.11	1.42	167
				10	94.66	2.80	79.79	19.35	0.86	248	95.78	0.51	87.73	11.23	1.04	146	98.99	1.89	83.04	16.41	0.55	173
				50	98.97	0.24	87.21	12.39	0.40	241	97.96	2.42	87.08	11.66	1.26	203	97.59	3.37	90.44	9.49	0.07	295
				100	97.73	4.78	81.87	17.26	0.87	206	99.13	2.6	87.65	11.38	0.97	221	96.36	4.98	88.18	11.61	0.21	190
				200	94.57	2.63	79.57	19.23	1.20	254	94.31	1.84	81.43	17.94	0.63	256	97.46	0.71	83.49	15.79	0.72	130
8	150	20	5	1	91.74	2.93	86.33	12.75	0.92	292	98.45	1.14	87.74	10.54	1.72	323	95.14	1.77	89.22	9.01	1.77	221
				5	95.40	4.33	79.13	20.19	0.68	369	96.79	3.69	88.53	10.14	1.33	168	97.11	2.93	84.91	15.02	0.07	242
				10	96.67	4.75	86.56	13.05	0.39	402	93.67	4.51	86.75	12.33	0.92	227	96.56	4.19	81.54	17.19	1.27	193
				50	94.68	4.90	86.6	12.95	0.45	301	93.89	2.24	81.39	16.67	1.94	292	95.32	4.01	81.06	17.61	1.33	255
				100	96.26	3.65	82.67	16.97	0.36	280	96.62	2.28	83.55	15.01	1.44	293	99.26	1	82.9	15.36	1.74	248
				200	99.09	3.39	75.41	23.34	1.25	405	98.45	4.52	84.36	13.89	1.75	164	98.82	3.81	84.14	15.02	0.84	129
				min	90.10	0.24	74.30	10.25	0.13	15.24	92.42	0.24	78.44	10.14	0.12	12.37	94.02	0.30	81.06	8.25	0.07	12.13
				max	99.73	4.90	89.16	24.99	1.99	410.19	99.90	4.67	89.02	19.82	2.01	406.72	99.96	4.99	91.51	17.61	1.99	348.28
				mean	95.41	3.17	81.83	17.26	0.91	194.55	96.30	2.45	83.88	15.08	1.05	174.61	96.99	2.73	86.18	12.85	0.97	149.97

I: No. of customers; J: No. of potential DCs; K: No. of capacity levels at each DC;
θ: Multiplier for the response time cost;
LAG: Lagrangean bound expressed as percentage of optimal solution; GAP: 100×(Heuristic Solution-Optimal Solution)/Optimal Solution;
SP: Computational time of the subproblems, MP: Computational time of the master problems, H: Computational time of the heuristics;
CPU: Total computational time in seconds (Note that SP, MP, and H are expressed as percentages of the total computational time, CPU).

Table 2.4: Comparison of the cutting plane method and the Lagrangean heuristic for high response time costs: MTO supply chain design (M/G/1 Case - CV = 1.5)

					Cutting Plane Method										Lagrangean Heuristic					
No.	I	J	K	θ	FC	VC	RC	TC	E(W)	$\bar{\rho}$	DC	CUT	ITR	CPU	LAG	GAP	SP	MAS	HEU	CPU
					(%)	(%)	(%)							(s)	(%)	(%)	(%)	(%)	(%)	(s)
1	50	10	3	400	39	39	23	212,766	9	0.48	7	35	6	11	97.20	2.85	80.66	17.73	1.61	53
				600	35	36	30	236,528	9	0.48	7	66	10	22	97.99	1.45	74.93	23.41	1.66	43
				800	41	30	29	255,872	7	0.37	9	54	7	12	94.97	1.59	74.03	24.22	1.75	36
				1000	38	28	33	274,163	7	0.37	9	1791	200	319	97.04	3.12	82.29	17.09	0.62	**127**
				1500	33	24	43	319,662	7	0.37	9	1791	200	220	96.45	0.91	78.66	20.88	0.46	**139**
				2000	29	21	50	365,025	7	0.37	9	54	177	624	98.88	2.29	83.5	15.01	1.49	**190**
2	50	20	3	400	47	23	30	228,976	19	0.51	13	78	7	1149	99.21	3.81	74.42	25.23	0.35	**152**
				600	47	20	33	262,325	16	0.44	15	90	7	1782	94.15	3.40	89.39	9.84	0.77	**323**
				800	43	18	39	291,067	16	0.44	15	90	7	1896	95.95	1.93	75.38	24.36	0.26	**372**
				1000	44	16	40	317,721	14	0.39	17	3383	200	1256	99.42	4.12	84.12	14.99	0.89	**380**
				1500	37	14	49	380,310	14	0.39	17	3383	200	1406	98.20	3.37	77.77	20.35	1.88	**390**
				2000	34	12	54	441,483	13	0.37	18	3582	200	1938	98.03	0.61	72.72	25.96	1.32	**430**
3	100	5	3	400	28	68	4	210,908	12	0.67	4	8	3	11	97.30	1.24	70.15	28.76	1.09	112
				600	29	66	5	215,230	9	0.61	4	8	3	12	98.66	1.52	83.42	15.54	1.04	143
				800	30	65	5	218,490	7	0.56	4	8	3	17	95.40	2.56	70.61	29.07	0.32	178
				1000	29	65	6	221,250	7	0.56	4	12	4	18	97.42	1.75	82.13	16.55	1.32	183
				1500	28	63	9	228,097	7	0.56	4	16	5	11	99.94	0.41	85.65	13.64	0.71	126
				2000	27	61	12	234,938	7	0.56	4	8	3	12	94.28	3.21	73.24	24.86	1.90	196
4	100	10	3	400	43	48	9	193,580	12	0.55	7	28	5	37	94.43	0.72	73.9	23.91	2.19	312
				600	44	46	9	201,406	8	0.48	7	7	2	36	95.41	1.49	78.58	21.19	0.23	318
				800	43	45	12	207,730	8	0.48	7	14	3	856	99.02	3.17	89.18	10.13	0.69	**426**
				1000	44	44	13	213,144	7	0.44	7	7	2	954	94.47	3.79	73.59	24.7	1.71	**435**
				1500	45	41	14	224,999	6	0.38	7	7	2	1538	96.88	0.22	80.55	18.48	0.97	**546**
				2000	44	39	17	235,026	5	0.36	7	14	3	2315	97.47	1.54	78.07	21.33	0.60	**422**
5	100	20	3	400	45	49	6	187,426	12	0.55	7	7	2	14	95.91	0.44	86.58	12.75	0.67	167
				600	43	48	9	193,341	12	0.55	7	28	5	38	99.11	1.58	70.43	28.76	0.81	215
				800	44	47	10	198,756	10	0.51	7	42	7	84	95.08	2.72	78.93	19.93	1.14	212
				1000	44	46	10	203,233	8	0.48	7	7	2	37	94.57	2.79	72.96	26.05	0.99	315
				1500	44	44	13	212,782	7	0.44	7	7	2	90	94.49	1.92	85.03	14.75	0.22	245
				2000	44	42	14	220,999	6	0.41	7	28	5	352	95.90	2.31	85.32	14.41	0.27	**223**
6	150	5	5	400	27	67	5	252,673	7	0.56	4	8	3	13	94.58	1.14	75.44	23.4	1.16	172
				600	27	66	8	259,205	7	0.56	4	4	2	77	99.96	1.69	85.11	13.26	1.63	187
				800	26	64	10	265,718	7	0.56	4	20	6	238	95.87	0.89	76.14	22.35	1.51	**219**
				1000	25	63	12	272,157	7	0.56	4	4	2	393	99.16	2.57	89.21	9.02	1.77	**211**
				1500	24	59	17	288,223	7	0.56	4	16	5	378	97.19	2.83	75.28	23.24	1.48	**267**
				2000	23	56	21	304,252	7	0.56	4	16	5	395	98.41	2.80	73.64	25.67	0.69	**345**
7	150	10	5	400	33	61	5	230,691	16	0.64	6	24	5	33	95.55	2.77	86.4	11.91	1.69	178
				600	34	60	6	235,314	12	0.59	6	18	4	126	95.65	1.26	80.57	18.09	1.34	**119**
				800	33	59	8	239,855	12	0.59	6	18	4	134	98.16	2.06	75.42	22.97	1.61	**121**
				1000	34	59	7	241,739	10	0.54	6	12	3	308	99.95	1.59	84.61	13.81	1.58	**249**
				1500	35	57	8	248,715	8	0.49	6	18	4	408	99.63	3.01	74.32	24.95	0.73	**298**
				2000	32	60	9	254,275	7	0.49	5	5	2	205	95.94	1.18	75.19	23.73	1.08	**145**
8	150	20	5	400	45	48	7	202,139	28	0.68	9	36	5	664	96.02	2.63	76.78	23.01	0.21	**445**
				600	46	46	8	208,164	23	0.63	9	18	3	1372	99.34	1.14	71.16	27.03	1.81	**647**
				800	45	49	6	212,702	14	0.56	8	16	3	2146	94.63	2.66	76.89	20.95	2.16	**546**
				1000	41	52	7	215,704	12	0.56	7	14	3	2009	95.62	3.41	81.59	18.04	0.37	**587**
				1500	42	50	8	222,842	10	0.52	7	7	2	2088	98.37	0.30	70.3	28.59	1.11	**458**
				2000	42	49	9	229,020	9	0.48	7	14	3	2595	95.82	3.53	85.84	12.93	1.23	**575**
				min	23	12	4	187,426	5	0.36	4	4	2	10.65	94.15	0.22	70.15	9.02	0.21	36
				max	47	68	54	441,483	28	0.68	18	3582	200	2595	99.96	4.12	89.39	29.07	2.19	647
				mean	37	47	16	245,638	10	0.51	7	311	28	638	97	2	79	20	1	282

I: No. of customers; J: No. of potential DCs; K: No. of capacity levels at each DC;
θ: Multiplier for the response time cost;
FC: Fixed cost; VC: Variable cost; RC: Response time cost; TC: Total cost;
$\mathbb{E}(W)$: Total expected waiting time;
$\bar{\rho}$: Average DC Utilization;
DC: Total no. of DCs selected open;
CUT: Number of cuts generated;
ITR: Number of iterations;
CPU(s): CPU time in sec.
LAG: Lagrangean bound expressed as percentage of optimal solution;
GAP: 100×(Heuristic Solution-Optimal Solution)/Optimal Solution;
SP: Computational times of the subproblems,
MP: Computational times of the master problems,
H: Computational times of the heuristics
(Note that SP, MP, and H are expressed as percentages of the total computational time, CPU).

Let $\mathbb{E}[W_j^n]$ denote the steady-state waiting time of a nth class customer at DC j. The problem is to simultaneously determine the location and the capacity of DCs $\mathbf{y}^* = \{y_{jk}^*\}$, and the allocation of multiclass customer demand to DCs $\mathbf{x}^* = \{x_{ij}^{n*}\}$ so as to minimize the cost of response time in addition to the sum of fixed cost of opening DCs and equipping them with sufficient processing capacity and the variable cost of serving customers.

Case I: Systems with Nonpreemptive Priorities

With nonpreemptive priorities (NPP), a customer being served cannot be ejected back into the queue if a higher priority customer enters the system. Therefore, once the DC has begun serving a customer, the service must be completed without interruption. For a DC j, modelled as an M/M/1 queue with multiple customer classes and NPP service discipline, the steady-state waiting time in the system (including the service time) $\mathbb{E}[W_j^n]$, for a customer of priority class n, is given by

$$\mathbb{E}[W_j^n] = \frac{1}{AB_{n-1}B_n} + \frac{1}{\mu_j}, \quad n = 1, 2, ..., N$$

where $A = \mu_j^2 / \sum_{n=1}^{N} \lambda^n$, $B_0 = 1$, and $B_n = 1 - \sum_{m=1}^{n} \lambda^m/\mu_j$, $\forall n \geq 1$.

In case of systems with two customer classes ($N = 2$, where h: high priority and l: low priority), the expression for $\mathbb{E}[W_j^n]$ reduces to:

$$\mathbb{E}[W_j^h] = \frac{\lambda_j^h}{\mu_j(\mu_j - \lambda_j^h)} + \frac{1}{\mu_j}, \quad \text{and} \quad \mathbb{E}[W_j^l] = \frac{\lambda_j^h + \lambda_j^l}{(\mu_j - \lambda_j^h)(\mu_j - \lambda_j^h - \lambda_j^l)} + \frac{1}{\mu_j}$$

The *total response time* is obtained by multiplying the steady-state expected waiting time by the expected arrival rates as follows:

$$\lambda_j^h \mathbb{E}[W_j^h] = \lambda_j^h \left(\frac{\lambda_j^h}{\mu_j(\mu_j - \lambda_j^h)} + \frac{1}{\mu_j} \right) = \frac{\lambda_j^h}{\mu_j - \lambda_j^h} \tag{2.23}$$

$$\lambda_j^l \mathbb{E}[W_j^l] = \lambda_j^l \left(\frac{\lambda_j^h + \lambda_j^l}{(\mu_j - \lambda_j^h)(\mu_j - \lambda_j^h - \lambda_j^l)} + \frac{1}{\mu_j} \right) \tag{2.24}$$

Note that the expression for the total response time of high priority class customers (2.23) is same as that of a single class customers and hence we can use the linearization scheme presented in Section 2.3. In order to linearize the expression for total response time for low priority class customers (2.24), we proceed as follows. Given that the function $\lambda_j^l \mathbb{E}[W_j^l]$ is convex, it can be approximated by infinite set of supporting hyperplanes that are tangent to $W_j^l(\lambda_j^l, \lambda_j^h, \mu_j^l)$ at various points $(\lambda_j^{lq}, \lambda_j^{hq}, \mu_j^{lq}), \forall q \in Q$, that is

$$W_j^l(.) = \max_{q \in Q} \left\{ W_j^{lq}(.) + (\lambda_j^l - \lambda_j^{lq}) \left(\frac{\partial W_j^{lq}(.)}{\partial \lambda_j^l} \right) + (\lambda_j^h - \lambda_j^{hq}) \left(\frac{\partial W_j^{lq}(.)}{\partial \lambda_j^h} \right) + (\mu_j - \mu_j^q) \left(\frac{\partial W_j^{lq}(.)}{\partial \mu_j} \right) \right\} \quad \forall j$$

which can be written as

$$W_j^l(.) \geq W_j^{lq}(.) + (\lambda_j^l - \lambda_j^{lq}) \left(\frac{\partial W_j^{lq}(.)}{\partial \lambda_j^l} \right) + (\lambda_j^h - \lambda_j^{hq}) \left(\frac{\partial W_j^{lq}(.)}{\partial \lambda_j^h} \right) + (\mu_j - \mu_j^q) \left(\frac{\partial W_j^{lq}(.)}{\partial \mu_j} \right) \quad \forall j, q \in Q$$

$$\text{or} \quad W_j^l(.) - \left(\frac{\partial W_j^{lq}(.)}{\partial \lambda_j^l} \right) \lambda_j^l - \left(\frac{\partial W_j^{lq}(.)}{\partial \lambda_j^h} \right) \lambda_j^h - \left(\frac{\partial W_j^{lq}(.)}{\partial \mu_j} \right) \mu_j \geq$$
$$W_j^{lq}(.) - \lambda_j^{lq} \frac{\partial W_j^{lq}(.)}{\partial \lambda_j^l} - \lambda_j^{hq} \frac{\partial W_j^{lq}(.)}{\partial \lambda_j^h} - \mu_j^q \frac{\partial W_j^{lq}(.)}{\partial \mu_j} \quad \forall j, q \in Q$$

where $\frac{\partial W_j^{lq}(.)}{\partial \lambda_j^l}$, $\frac{\partial W_j^{lq}(.)}{\partial \lambda_j^h}$ and $\frac{\partial W_j^{lq}(.)}{\partial \mu_j}$ are the subgradients at points $q \in Q$, and can be computed by taking the partial derivatives of the expression (2.24).

The resulting linear MIP model $[P_{(Q^q)}]$ is as follows:

$$\min \quad \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk}y_{jk} + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=h,l} c_{ij}\lambda_i^n x_{ij}^n + \sum_{j=1}^{J}(t_h R_j + t_l W_j^l) \tag{2.25}$$

$$\text{s.t.} \quad \sum_{i=1}^{I}\sum_{n=l,h} \lambda_i^n x_{ij}^n - \sum_{k=1}^{K} \mu_{jk} z_{jk} = 0 \qquad \forall j \tag{2.26}$$

$$\sum_{k=1}^{K} y_{jk} \leq 1 \qquad \forall j \tag{2.27}$$

$$\sum_{j=1}^{J} x_{ij}^n = 1 \qquad \forall i, n \tag{2.28}$$

$$z_{jk} - y_{jk} \leq 0 \qquad \forall j, k \tag{2.29}$$

$$\rho_j^h - \frac{1}{(1+R_j^q)^2} R_j \leq \frac{(R_j^q)^2}{(1+R_j^q)^2} \qquad \forall j, q \in Q \tag{2.30}$$

$$\rho_j^h - \sum_{k=1}^{K} z_{jk} = 0 \qquad \forall j \tag{2.31}$$

$$\begin{aligned} W_j^l - \left(\frac{\partial W_j^{lq}(.)}{\partial \lambda_j^l}\right)\lambda_j^l - \left(\frac{\partial W_j^{lq}(.)}{\partial \lambda_j^h}\right)\lambda_j^h - \left(\frac{\partial W_j^{lq}(.)}{\partial \mu_j}\right)\mu_j \geq \\ W_j^{lq}(.) - \lambda_j^{lq}\frac{\partial W_j^{lq}(.)}{\partial \lambda_j^l} - \lambda_j^{hq}\frac{\partial W_j^{lq}(.)}{\partial \lambda_j^h} - \mu_j^q \frac{\partial W_j^{lq}(.)}{\partial \mu_j} \qquad \forall j, q \in Q \end{aligned} \tag{2.32}$$

$$y_{jk} \in \{0,1\}; \quad 0 \leq x_{ij}^n, z_{jk} \leq 1; \quad \rho_j, R_j, W_j^l \geq 0 \qquad \forall i, j, k, n \tag{2.33}$$

The model with large number of constraints in (2.30) and (2.32) is amenable to the cutting plane method, where one can start with an initial subset of constraints, and others are included as needed. For an initial and finite set of points $\{x^q, y^q\}_{Q^q \subset Q}$ considered, $[P_{(Q^q)}]$ is a relaxation of the full problem $[P_{(Q)}]$, hence a lower bound to $[P_{G(H)}]$ or $[P_G]$ is provided by the optimal objective function value $v(P_{G(H^q)})$, where the lower and upper bounds are given by:

$$Z_{LB}(\overline{x},\overline{y}) = v(P_{(Q^q)}) = \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk}\overline{y}_{jk} + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=h,l} c_{ij}\lambda_i^n \overline{x}_{ij}^n + \sum_{j=1}^{J}(t_h \overline{R}_j + t_l \overline{W}_j^l)$$

$$Z_{UB}(\overline{x},\overline{y}) = \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk}\overline{y}_{jk} + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=h,l} c_{ij}\lambda_i^n \overline{x}_{ij}^n + \sum_{j=1}^{J}\sum_{n=h,l} t_n \lambda_j^n W_j^n\,(\overline{x},\overline{y})$$

where,

$$W_j^h(.) = \frac{\lambda_j^h}{\mu_j - \lambda_j^h} \quad \text{and} \quad W_j^l(.) = \lambda_j^l \left(\frac{\lambda_j^h + \lambda_j^l}{(\mu_j - \lambda_j^h)(\mu_j - \lambda_j^h - \lambda_j^l)} + \frac{1}{\mu_j} \right)$$

If the upper bound coincides with the lower bound then $(\overline{x}, \overline{y})$ is an optimal solution to $[P]$, and the procedure is terminated. If not, then using the current solution $(\overline{x}, \overline{y})$, we get the estimates of waiting time $W_j^h(.)$ and $W_j^l(.)$, and its subgradients $\frac{\partial W_j^{q_{new}}}{\partial \lambda_j}$ and $\frac{\partial W_j^{q_{new}}}{\partial \mu_j}$, where $\lambda_j^{q_{new}} = \sum_{i=1}^{I} \sum_{l=1}^{L} \lambda_{il} \overline{x}_{ijl}^q$ and $\mu_j^{q_{new}} = \sum_{k=1}^{K} \mu_{jk} \overline{y}_{jk}^q$ and generate two cuts of the form

$$\rho_j^h - \frac{1}{(1+R_j^q)^2} R_j \leq \frac{(R_j^q)^2}{(1+R_j^q)^2} \qquad \forall j, q \in Q \quad (2.34)$$

$$\begin{aligned} & W_j^h - \left(\frac{\partial W_j^{q_{new}}}{\partial \lambda_j} \right) \sum_{i=1}^{I} \sum_{l=1}^{L} \lambda_{il} x_{ijl} - \left(\frac{\partial W_j^{q_{new}}}{\partial \mu_j} \right) \sum_{k=1}^{K} \mu_{jk} y_{jk} \geq \\ & \quad W_j \left(\sum_{i=1}^{I} \sum_{l=1}^{L} \lambda_{il} \overline{x}_{ijl}^q, \sum_{k=1}^{K} \mu_{jk} \overline{y}_{jk}^q \right) - \left(\frac{\partial W_j^{q_{new}}}{\partial \lambda_j} \right) \sum_{i=1}^{I} \sum_{l=1}^{L} \lambda_{il} \overline{x}_{ijl}^q - \left(\frac{\partial W_j^{q_{new}}}{\partial \mu_j} \right) \sum_{k=1}^{K} \mu_{jk} \overline{y}_{jk}^q \quad \forall j \end{aligned} \quad (2.35)$$

This new set of constraints is appended to constraint set (2.30) and (2.32) respectively and the procedure is repeated again.

In settings where the assumption of exponential service times is too restrictive, the model can be extended to deal with multiple customer classes with general service times distributions, for which the steady-state waiting time is given by:

$$\mathbb{E}[W_j^n] = \frac{\sum_{n=1}^{N} \lambda_j^n (1 + \sigma_j^2 \mu_j^2)}{2\mu_j \prod_{m=n-1}^{n} (\mu_j - \lambda_j^1 - \ldots - \lambda_j^m)} + \frac{1}{\mu_j}, \quad n = 1, 2, \ldots, N$$

For two priority class with nonpreemptive priorities, these expressions reduce to

$$\lambda_j^h \mathbb{E}[W_j^h] = \lambda_j^h \left\{ \frac{\lambda_j^h \left(1 + \sigma_j^2 \mu_j^2\right)}{2\mu_j (\mu_j - \lambda_j^h)} + \frac{1}{\mu_j} \right\} \quad (2.36)$$

$$\lambda_j^l \mathbb{E}[W_j^l] = \lambda_j^l \left\{ \frac{\left(\lambda_j^h + \lambda_j^l\right) \left(1 + \sigma_j^2 \mu_j^2\right)}{2(\mu_j - \lambda_j^h)(\mu_j - \lambda_j^h - \lambda_j^l)} + \frac{1}{\mu_j} \right\} \quad (2.37)$$

Case II: Systems with Preemptive Priorities

In case of systems where preempting (interrupting) the service of customers with lower response time costs to allow customers with higher response time costs to be served immediately is permitted, the steady-state waiting time is given by:

$$\mathbb{E}[W_j^n] = \frac{1/\mu_j}{B_{n-1}B_n}, \quad n = 1, 2, ..., N$$

where $B_0 = 1$ and $B_n = 1 - \sum_{m=1}^{n} \lambda^m/\mu_j$, $\forall n \geq 1$.

The total response times are given by:

$$\lambda_j^h \mathbb{E}[W_j^h] = \frac{\lambda_j^h}{\mu_j - \lambda_j^h} \quad \text{and} \quad \lambda_j^l \mathbb{E}[W_j^l] = \frac{\lambda_j^l \mu_j}{(\mu_j - \lambda_j^h)(\mu_j - \lambda_j^h - \lambda_j^l)}$$

These expressions can be linearized using the procedure shown above.

2.9.2 Systems with General Demand Processes and Service Time Distributions

In case of systems with general arrival processes and service time distributions, where the demand forms independent renewal processes, DCs can be modelled as GI/G/1 queues. However, two technical difficulties must be addressed: (i) The superposition of renewal processes does not necessarily yield a renewal process, and therefore, the arrival process to each facility may not be a renewal process, (ii) There are no exact expressions for the expected waiting time in a GI/G/1 queue. As pointed out by Benjaafar et al. [21, 22] the first difficulty can be handled by approximating superposed renewal processes by a renewal process whose coefficient of variation is obtained via a two-moment approximation, such as the asymptotic method described in Albin [6] and Whitt [152]. The second difficulty can be addressed by using one of the reasonably approximations for the expected waiting time in a GI/G/1 queue (see Buzacott and Shanthikumar [33] and Wolff [154]). One such approximation for the expected waiting time (service plus queuing time) is due to Whitt [152]:

$$\begin{aligned} W_j(GI/G/1) &\simeq \left(\frac{c_a^2 + c_s^2}{2}\right)\left(\frac{\tau_j \rho_j}{1-\rho_j}\right) + \tau_j \\ &\simeq \left(\frac{\beta_j^2 \lambda_j^2 + \sigma_j^2 \mu_j^2}{2}\right) \frac{\lambda_j}{\mu_j(\mu_j - \lambda_j)} + \frac{1}{\mu_j} \qquad \forall j \end{aligned}$$

where τ_j represent the mean service time at DC j ($\tau_j = 1/\mu_j$), ρ_j be the utilization of DC j ($\rho_j = \lambda_j/\mu_j$), c_s^2 be the squared coefficient of variation of service times ($c_s^2 = \sigma_j^2/\tau_j^2 = \sigma_j^2\mu_j^2$), β_j^2 be the variance of arrival times at DC j, and c_a^2 be the squared coefficient of variation of arrival times ($c_a^2 = \beta_j^2\lambda_j^2$). It can be shown that the total expected waiting time for a GI/G/1 queue ($\lambda_j W_j(GI/G/1)$) is convex in the arrival rate [66]. Alternatively, one may focus on heavy traffic regimes ($\rho \to 1$) for which explicit results for GI/G/1 queue are available. Some of these results can be found in Peterson [105]. Along with these approximate expressions and explicit results, the cutting plane method presented above can be used to tackle these cases.

2.10 Concluding Remark

In this chapter, we modelled and analyzed the effect of response time consideration on the design of MTO supply chain networks. We presented an MTO supply chain design model that captures the trade-off among response time, the fixed cost of opening DCs and equipping them with sufficient capacity, and the transportation cost associated with serving customers. Under the assumption that the customer demand follows Poisson process and service times follow general distribution, the DCs were modelled as a network of single-server queues, whose capacity levels and locations are decision variables. We presented a non-linear MIP formulation, a linearization procedure, a cutting plane method, and a Lagrangean heuristic. Our computational results indicate that while the cutting plane method provides optimal solution for moderate instances of the problem in few iterations, the Lagrangean heuristic provides solution that is within 5% of the optimal for the test instances in reasonable computation times. We used the models to demonstrate empirically that substantial improvement (decrease) in response time can be achieved with a minimal increase in total cost associated with designing supply chains. Also we showed that the supply chain configuration (DC location and capacity, and allocation of customers to DCs) obtained using the model that considers congestion can be very different from those obtained using the traditional models that ignores response time. Furthermore, the inclusion of response time in the objective function may not satisfy the closest-assignment property and causes splitting of demand in few cases. We showed that our solution procedures can be easily extended to deal to MTO systems with multiple customer classes and general demand processes and service time distributions.

Chapter 3

Response Time Reduction in ATO Supply Chain Design[†]

3.1 Introduction

ATO system is a business strategy adopted by many firms to meet the dual needs of mass customization and shorter response time (e.g. Dell, IBM, Gateway, National Bicycle, Nike) [133]. ATO systems facilitate delayed product differentiation by maintaining inventories of semi-finished products/sub-assemblies while delaying the final assembly of finished products until the customer orders arrive. Maintaining inventory of common semi-finished products allows firms to aggregate demands across different finished products, thereby reducing safety stock inventories (due to risk pooling effects). Furthermore, it increases firm's responsiveness to cater to unpredictable changes in the demand mix and reduces the response time. However, the time to assemble a finished product following an order depends on the assembly capacity, workload allocated, and the variation in processing times of the product mix. Inadequate assembly capacity, suboptimal allocation of customer demand, coupled with variability in processing times of product mix can spell long response times, which could easily wipe out firm's profit margin and diminish its competitive edge.

[†] A version of the materials in this chapter has appeared in the paper: N. K. Vidyarthi, S. Elhedhli, and E. M. Jewkes (2009), Response time reduction in make-to-order and assemble-to-order supply chain design, *IIE Transactions*, 41(5), 448-466.

The objective of this chapter is to model the effect of congestion on response time in the design of a two-echelon ATO supply chain, where a set of plants and DCs are to be established to assemble and distribute finished products to a set of customers with stochastic demand. We formulate a nonlinear MIP model of the problem. We present a Lagrangean relaxation of the model that exploits the echelon structure of the problem to decompose into two problems - one relates to MTS echelon and other relates to MTO echelon. We propose a heuristic to construct feasible solution. To the best of our knowledge, this is one of the early attempts that explicitly account for congestion in the design of two-echelon supply chain network design.

The remainder of this chapter is organized as follows. In Section 3.2, we review related literature. Section 3.3 presents a nonlinear MIP model of the ATO supply chain design problem. Section 3.4 presents a Lagrangean relaxation of the model and the heuristic. Computational results are reported in Section 3.5. Finally, in Section 3.6, we conclude with some remarks.

3.2 Related Literature

The significance of offering a high level of product variety to customers, while maintaining reasonable response times and costs has prompted a number of researchers to address issues related to ATO supply chain. We refer the reader to Song and Zipkin [133] for reviews of the ATO literature. Most of the papers on ATO systems address tactical/operational planning issues such as optimal inventory levels, service level, component commonality etc. assuming that the facility location and capacity decisions are fixed. It is worthwhile noting that the strategic configuration of the supply chain has a long lasting impact on the firm and influences its tactical level decisions. It can be construed from the above discussion that, despite such impressive anecdotal evidence of the importance of stochastic demand and response time, few analytical model have examined strategic design of supply chain from a response time perspective.

Another stream of related literature is on *supply chain network design*. We refer the reader to Section 2.2.2 of the previous chapter for a brief review of related literature on supply chain network design. Most of the earlier work on supply chain network design are in deterministic settings. As pointed out in the previous chapter, to the best of our knowledge, Huang et al. [75] is the only paper to model the effect of congestion in the design of distribution networks. The originality of our work is the explicit modelling of congestion using a queueing framework in the design of two-echelon ATO supply chains. We also present solution approaches to deal such highly nonlinear large-scale models.

3.3 Model Formulation

We consider the problem of designing an ATO supply chain, where we seek to locate a set of plants and DCs to distribute a product with non-trivial bill-of-material to a set of customers with stochastic demand. The DCs will act as intermediate facilities between the plants and the customers and facilitate the shipment of products between the two echelons, as shown in Figure 3.1.

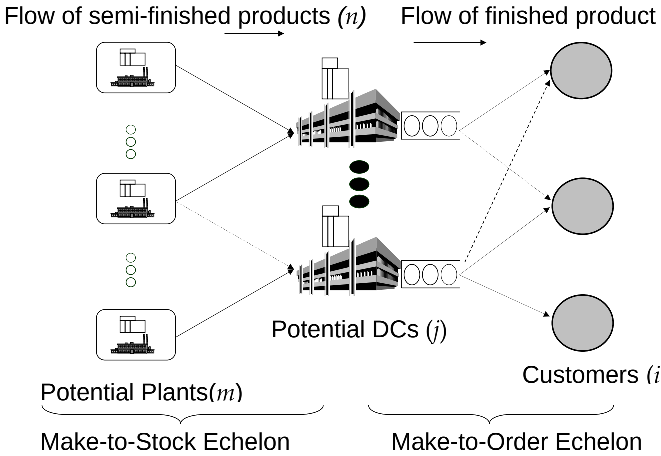


Figure 3.1: An assemble-to-order supply chain network

The semifinished products are produced at the plants and shipped to the DCs, where an inventory of semi-finished products is maintained. Once the demand is realized at the customers' end, the order is placed to the DC and the final product is assembled and the demand is met. Hence, the supply chain network is a combination of make-to-stock echelon (plant-DC echelon) and the make-to-order echelon (DC-customer echelon). The problem environment is characterized by stochastic customer demand that has to be satisfied from a set of DCs where sufficient capacity has to be acquired in order to avoid long response times. To model this problem, we define the following additional notation:

m : Index for potential plants, $m = 1, 2, \ldots, M$.

n : Index for semi-finished products, $n = 1, 2, \ldots, N$.

g_m : Fixed cost of opening a plant at location m (\$/period).

p_m : Maximum available capacity of plant m (units).

c'_{jmn} : Unit production and transportation cost for semi-finished product n from plant m to DC j.

η_n : Number of units of semi-finished product n required to make one unit of finished product.

u_m : Decision variable that equals 1, if plant m is opened; 0, otherwise.

v_{jmn} : Number of units of semi-finished product n produced at plant m and shipped to DC j.

Under the assumption that the demand at customer i is an independent random variable that follows a Poisson process with mean λ_i and the service time at each DC follows a general distribution, each DC is modelled as M/G/1 queue, whose mean service rate, if it is allocated capacity level k, is given by $\mu_j = \sum_{k=1}^{K} \mu_{jk} y_{jk}$ and the variance in service times is given by $\sigma_j^2 = \sum_{k=1}^{K} \sigma_{jk}^2 y_{jk}$. Under steady state conditions ($\lambda_j < \mu_j$) and FCFS queuing discipline, the *total average waiting time* for the entire system (service plus queuing time) is given by (2.1). The resulting non-linear MIP formulation that simultaneously determines the location and capacity of plants and DCs, the shipment levels from plants to DCs, and allocation of customers to DCs by minimizing response time costs in addition to fixed cost of location and capacity acquisition, and the variable cost of production and transportation costs between echelons is as follows:

$$
\begin{aligned}
[P_{ATO}] : \min \quad & \sum_{m=1}^{M} g_m u_m + \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} y_{jk} + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N} c'_{jmn} v_{jmn} + \sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}\lambda_i x_{ij} + \\
& \frac{t}{2}\sum_{j=1}^{J}\left(1+\sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk}y_{jk} - \sum_{i=1}^{I}\lambda_i x_{ij}} + \\
& \frac{t}{2}\sum_{j=1}^{J}\left(1-\sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk}y_{jk}} \qquad (3.1)
\end{aligned}
$$

$$
\begin{aligned}
\text{s.t.} \quad & \sum_{j=1}^{J}\sum_{n=1}^{N} v_{jmn} \le p_m u_m && \forall m && (3.2)\\
& \sum_{m=1}^{M} v_{jmn} = \sum_{i=1}^{I} \eta_n \lambda_i x_{ij} && \forall j, n && (3.3)\\
& \sum_{i=1}^{I} \lambda_i x_{ij} \le \sum_{k=1}^{K} \mu_{jk} y_{jk} && \forall j && (3.4)\\
& \sum_{k=1}^{K} y_{jk} \le 1 && \forall j && (3.5)\\
& \sum_{j=1}^{J} x_{ij} = 1 && \forall i && (3.6)\\
& 0 \le x_{ij} \le 1, \quad y_{jk}, u_m \in \{0,1\}, \quad v_{jmn} \ge 0 && \forall i, j, k, m, n && (3.7)
\end{aligned}
$$

The objective function (3.1) consists of fixed cost of opening plants, fixed cost of locating DCs and equipping them with the required capacity level, the variable cost of producing and procuring semi-finished product, the variable cost of serving customers from DCs, and the total waiting time costs at the DCs. Constraints (3.2) are capacity restrictions on the opened plants and permit the use of opened plants only. Constraints (3.3) are commodity flow conservation equations at the DCs. Constraints (3.4) ensure that the steady state conditions at the DCs are met. Constraints (3.5) ensure that at most one capacity level is selected at a DC whereas constraints (3.6) ensure that the total demand is met. Constraints (3.7) are nonnegativity and binary constraints. The model can be extended to include the safety stock inventory costs at DCs as discussed in Buzacott and Shanthikumar [33]. Also refer to Benjaafar et al. [22].

Model $[P_{ATO}]$ can be linearized using a simple transformation and the piecewise linear approximation presented in the previous chapter. This would yield a linear model with large number of constraints that can be solved to optimality using the cutting plane method (presented in Section 3.4). However, our initial computational testing shows that such an approach will yield excessive runtimes, even for small size problems (primarily due to the large number of constraints and variables). This motivates the development of the Lagrangean heuristic.

3.4 Lagrangean Relaxation

There are number of ways in which the model can be relaxed in Lagrangean fashion (refer to Klose and Drexl[81] and references therein). In this paper, we exploit the echelon structure of the ATO supply chain using Lagrangean relaxation to decompose the model into two subproblems. Note that in $[P_{ATO}]$, constraints (3.2) relate to the MTS echelon, constraints (3.4)-(3.6) relate to the MTO echelon, whereas constraints (3.3) are the flow conservation constraints that link the two echelons. Upon relaxing the flow conservation constraints (3.3) with dual multipliers β_{jn}, the problem decomposes into two subproblems:

$$[SP_{MTS}] : \min \quad \sum_{m=1}^{M} g_m u_m + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N}(c'_{jmn} - \beta_{jn})v_{jmn} \tag{3.8}$$

$$\text{s.t.} \quad \sum_{j=1}^{J}\sum_{n=1}^{N} v_{jmn} \leq p_m u_m \qquad \forall m \tag{3.9}$$

$$u_m \in \{0,1\}, \quad v_{jmn} \geq 0 \qquad \forall j, m, n \tag{3.10}$$

$$
\begin{aligned}
[SP_{MTO}] : \min \quad & \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk}y_{jk} + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=1}^{N}(c_{ij} + \eta_n\beta_{jn})\lambda_i x_{ij} + \\
& \frac{t}{2}\sum_{j=1}^{J}\left(1 + \sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk}y_{jk} - \sum_{i=1}^{I}\lambda_i x_{ij}} + \\
& \frac{t}{2}\sum_{j=1}^{J}\left(1 - \sum_{k=1}^{K} CV_{jk}^2 y_{jk}\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk}y_{jk}} && (3.11)\\
\text{s.t.} \quad & \sum_{i=1}^{I}\lambda_i x_{ij} \le \sum_{k=1}^{K}\mu_{jk}y_{jk} && (3.12)\\
& \sum_{j=1}^{J} x_{ij} = 1 \qquad \forall i && (3.13)\\
& \sum_{k=1}^{K} y_{jk} \le 1 \qquad \forall j && (3.14)\\
& 0 \le x_{ij} \le 1, \quad y_{jk} \in \{0,1\} \qquad \forall i,j,k && (3.15)
\end{aligned}
$$

Subproblem $[SP_{MTS}]$ is a linear MIP model that determines the location of plants and the flow of semi-finished products into the DCs, whereas the subproblem $[SP_{MTO}]$ is a non-linear MIP model that provides the location and capacity level of DCs and the allocation of customers to DCs. Note that the subproblem $[SP_{MTO}]$ is the MTO supply chain design model presented in the previous chapter and hence we use the proposed cutting plane algorithm to solve it. From model $[P_{ATO}]$, we can derive some valid constraints:

$$
\begin{aligned}
& \sum_{m=1}^{M} p_m u_m \ge \sum_{i=1}^{I}\lambda_i && (3.16)\\
& \sum_{m=1}^{M} v_{jmn} \le \eta_n\left(\max_k \mu_{jk}\right) \qquad \forall j,n && (3.17)\\
& \sum_{j=1}^{J}\sum_{m=1}^{M} v_{jmn} \ge \eta_n\sum_{i=1}^{I}\lambda_i \qquad \forall n && (3.18)
\end{aligned}
$$

Constraints (3.16) are aggregate capacity constraints for the MTS echelon. Constraints (3.17) is derived from (3.3) and (3.4) and constraints (3.18) follows from (3.3) and (3.6). Constraints (3.17) imply that the total flow of semi-finished products through a DC should not exceed the DC's maximum throughput capacity, whereas constraints (3.18) ensure that the flow of every semi-finished product from plants to DC is at least equal to the bill-of-material times the demand of that product from all the customers. These constraints are redundant in the original MIP formulation, but they improve the quality of the subproblem solutions in terms of the feasibility to the original problem upon relaxing the flow conservation constraints. This results in better heuristic solutions. Therefore, we add these set of constraints to $[SP_{MTS}]$ as follows:

$$
\begin{aligned}
[SP_{MTS}]: \quad \min \quad & \sum_{m=1}^{M} g_m u_m + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N}(c'_{jmn} - \beta_{jn})v_{jmn} && && (3.19)\\
\text{s.t.} \quad & \sum_{j=1}^{J}\sum_{n=1}^{N} v_{jmn} \leq p_m u_m && \forall m && (3.20)\\
& \sum_{m=1}^{M} P_m u_m \geq \sum_{i=1}^{I} \lambda_i && && (3.21)\\
& \sum_{m=1}^{M} v_{jmn} \leq \eta_n \left(\max_k \mu_{jk}\right) && \forall j, n && (3.22)\\
& \sum_{j=1}^{J}\sum_{m=1}^{M} v_{jmn} \geq \eta_n \sum_{i=1}^{I} \lambda_i && \forall n && (3.23)\\
& u_m \in \{0,1\}, \quad v_{jmn} \geq 0 && \forall j, m, n && (3.24)
\end{aligned}
$$

3.4.1 The Lower Bound

The Lagrangean lower bound is given by the solution of the Lagrangean dual problem, $\max_\beta\{v(SP_{MTS}) + v(SP_{MTO})\}$ which is equivalent to:

$$
\begin{aligned}
\max_\beta\{ & \min_{h \in I_{u,v}} \sum_{m=1}^{M} g_m u_m^h + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N}(c'_{jmn} - \beta_{jn})v_{jmn}^h + \\
& \min_{h \in I_{x,y}} \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} y_{jk}^h + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=1}^{N}(c_{ij} + \eta_n \beta_{jn})\lambda_i x_{ij}^h + \\
& \frac{t}{2}\sum_{j=1}^{J}\left(1 + \sum_{k=1}^{K} CV_{jk}^2 y_{jk}^h\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}^h}{\sum_{k=1}^{K}\mu_{jk} y_{jk}^h - \sum_{i=1}^{I}\lambda_i x_{ij}^h} + \\
& \frac{t}{2}\sum_{j=1}^{J}\left(1 - \sum_{k=1}^{K} CV_{jk}^2 y_{jk}^h\right)\frac{\sum_{i=1}^{I}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk} y_{jk}}\}.
\end{aligned}
$$

This can be explicitly written as:

$$
\begin{aligned}
[MP] : \max_{\beta} \quad & \theta_1 + \theta_2 \\
\text{s.t.} \quad & \theta_1 + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N} v_{jmn}^{h}\beta_{jn} \leq \sum_{m=1}^{M} g_m u_m^h + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N} c'_{jmn} v_{jmn}^{h} && h \in I_{u,v} \\
& \theta_2 - \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=1}^{N} (\eta_n \lambda_i x_{ij}^h)\beta_{jn} \leq \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} y_{jk}^h + \sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}\lambda_i x_{ij}^h + \\
& \qquad \frac{t}{2}\sum_{j=1}^{J}\left(1 + \sum_{k=1}^{K} CV_{jk}^2 y_{jk}^h\right) \frac{\sum_{i=1}^{I}\lambda_i x_{ij}^h}{\sum_{k=1}^{K}\mu_{jk} y_{jk}^h - \sum_{i=1}^{I}\lambda_i x_{ij}^h} + \\
& \qquad \frac{t}{2}\sum_{j=1}^{J}\left(1 - \sum_{k=1}^{K} CV_{jk}^2 y_{jk}^h\right) \frac{\sum_{i=1}^{I}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk} y_{jk}} && h \in I_{x,y}
\end{aligned}
$$

where $I_{u,v}$ is the index set of feasible points of the set:

$$\{(u_m, v_{jmn}) : (40) - (43); \quad u_m \in \{0,1\}; \quad v_{jmn} \geq 0, \quad \forall j, m, n\}$$

and $I_{x,y}$ is the index set of feasible points of the set:

$$\{(x_{ij}, y_{jk}) : (32) - (34); \quad x_{ij} \geq 0; \quad y_{jk} \in \{0,1\}, \quad \forall i, j, k\}.$$

We use Kelley's cutting plane method [78], in which the point $\bar{\beta}$ is the solution of the relaxed master problem $[RMP]$, defined on subsets $\bar{I}_{u,v} \subset I_{u,v}$ and $\bar{I}_{x,y} \subset I_{x,y}$. This $\bar{\beta}$ from $[RMP]$ is used to solve the subproblems $[SP_{MTS}]$ and $[SP_{MTO}]$, and generate two cuts of the form:

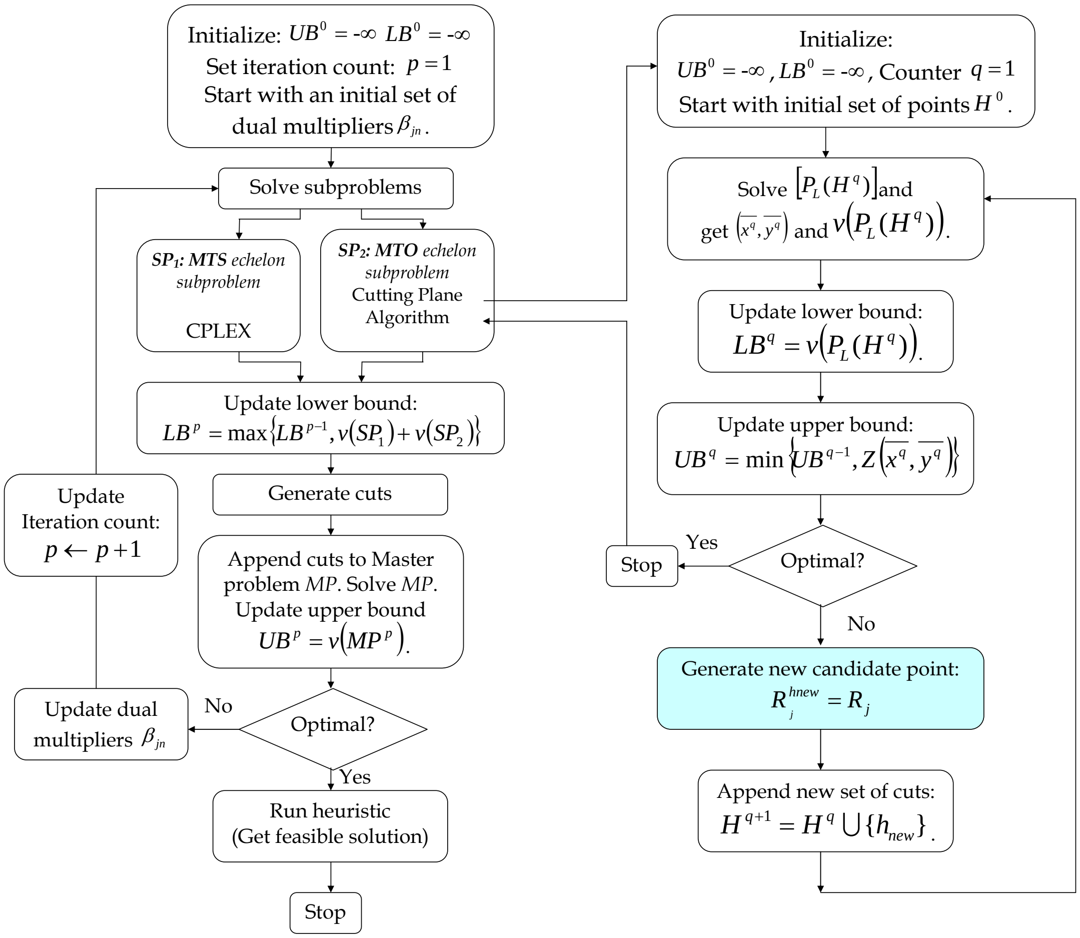
$$\theta_1 + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N} v_{jmn}^{\bar{i}}\beta_{jn} \leq \sum_{m=1}^{M} g_m u_m^{\bar{i}} + \sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{n=1}^{N} c'_{jmn} v_{jmn}^{\bar{i}} \tag{3.25}$$

$$
\begin{aligned}
& \theta_2 - \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=1}^{N} (\eta_n \lambda_i x_{ij}^{\bar{i}'})\beta_{jn} \leq \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk} y_{jk}^{\bar{i}'} + \sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}\lambda_i x_{ij}^{\bar{i}'} + \\
& \quad \frac{t}{2}\sum_{j=1}^{J}\left(1 + \sum_{k=1}^{K} CV_{jk}^2 y_{jk}^{\bar{i}'}\right) \frac{\sum_{i=1}^{I}\lambda_i x_{ij}^{\bar{i}'}}{\sum_{k=1}^{K}\mu_{jk} y_{jk}^{\bar{i}'} - \sum_{i=1}^{I}\lambda_i x_{ij}^{\bar{i}'}} + \frac{t}{2}\sum_{j=1}^{J}\left(1 - \sum_{k=1}^{K} CV_{jk}^2 y_{jk}^{\bar{i}'}\right) \frac{\sum_{i=1}^{I}\lambda_i x_{ij}^{\bar{i}'}}{\sum_{k=1}^{K}\mu_{jk} y_{jk}^{\bar{i}'}}
\end{aligned} \tag{3.26}
$$

The index sets $\bar{I}_{u,v}$ and $\bar{I}_{x,y}$ are updated as $\bar{I}_{u,v} \cup \{\bar{i}\}$ and $\bar{I}_{x,y} \cup \{\bar{i}'\}$, respectively, as the algorithm proceeds through the iterations.

3.4.2 The Heuristic: Finding a Feasible Solution

The first subproblem $[SP_{MTS}]$ provides the location of plants (u_m) and the flow of semifinished into the DCs (v_{jmn}), whereas the second subproblem $[SP_{MTO}]$ provides the location and the capacity decisions of the DCs (y_{jk}), the assignment of customers to DCs (x_{ij}). Note that the link between the two subproblems is the flow balance of products in and out of DCs. Hence, a feasible solution to problem $[P_{ATO}]$ can be constructed by solving $[SP_{MTS}]$ with the additional set of constraints $\sum_{m=1}^{M} v_{jmn} = \sum_{i=1}^{I} \eta_n \lambda_i \overline{x}_{ij}$, where $\overline{x}_{ij}$ is obtained from the solution of $[SP_{MTO}]$. The overall procedure is shown in Figure 4.2. Computational results are provided next.



Initialize: $UB^0 = -\infty$ $LB^0 = -\infty$
Set iteration count: $p = 1$
Start with an initial set of dual multipliers β_{jn}.
Solve subproblems
SP1: MTS echelon subproblem
CPLEX
SP2: MTO echelon subproblem
Cutting Plane Algorithm
Update lower bound: $LB^p = \max\{LB^{p-1}, v(SP_1) + v(SP_2)\}$
Generate cuts
Append cuts to Master problem MP. Solve MP. Update upper bound $UB^p = v(MP^p)$.
Optimal?
No
Yes
Update dual multipliers β_{jn}
Update Iteration count: $p \leftarrow p+1$
Run heuristic (Get feasible solution)
Stop
Initialize: $UB^0 = -\infty$, $LB^0 = -\infty$, Counter $q = 1$
Start with initial set of points H^0.
Solve $[P_L(H^q)]$ and get $(\overline{x^q}, \overline{y^q})$ and $v(P_L(H^q))$.
Update lower bound: $LB^q = v(P_L(H^q))$.
Update upper bound: $UB^q = \min\{UB^{q-1}, Z(\overline{x^q}, \overline{y^q})\}$
Optimal?
Yes
Stop
No
Generate new candidate point: $R_j^{hnew} = R_j$
Append new set of cuts: $H^{q+1} = H^q \cup \{h_{new}\}$.

Figure 3.2: Solution procedure for two-echelon ATO supply chain design

3.5 Computational Results and Insights

The proposed solution procedure was coded in C and the MIP problems were solved using ILOG CPLEX 10.1 Callable Library on a Sun Blade 2500 workstation with 1.6-GHz UltraSPARC IIIi processors. The test instances are derived from the 2000 census data consisting of 150 largest cities in the continental United States (as done in [48]. We generate eight sets of test problems by setting the number of customers (I) to the 50, 100, and 150 largest cities, and the potential DC locations (J) to the 5, 10, and 20 most populated cities. The demand rates λ_i are obtained by dividing the population of those cities by 10^3. The unit transportation costs c_{ij} are obtained by dividing the great-circle distance between the customer i and the potential DC location j by 100. The service rate of DC j equipped with capacity level k, is set to $\mu_{jk} = \beta_k \sum_i \lambda_i$ (where $\beta_k = 0.15, 0.20, 0.45$ for $I = 50$; $\beta_k = 0.10, 0.20, 0.30$ for $I = 100$; $\beta_k = 0.10, 0.15, 0.20, 0.30, 0.45$ for $I = 150$). The fixed costs of the DC, f_{jk} are set to $100 \times \sqrt{\mu}_{jk}$. The capacities of plants are set to: $P_m = U[0.1, 0.5] \times \sum_{i=1}^{I} \lambda_i$ whereas their fixed cost are obtained using: $g_m = U[1000, 2000] \times \sqrt{(}p_m)$. For the M/G/1, the coefficient of variation (CV) to set to 1.5. The average response time cost t is set to $\theta \times \frac{\sum_i \sum_j (\lambda_i c_{ij})}{I \times J}$, where θ is the response time cost multiplier and $\lambda_i c_{ij}$ denotes the total production and transportation cost associated with the order from i^{th} customer served by j^{th} DC. The response time costs is varied by changing the multiplier θ to 0.1, 1, 5, 10, 50, 100, and 200. The production coefficients (bill-of-material) η_n were randomly generated in the range $U[1, 5]$ and rounded up to the nearest integer value.

3.5.1 Insights

Incorporating response time in the design of two-echelon supply chains impacts the location and capacity of DCs and the allocation of customer demand to DCs. This also effects the location and capacity of plants and inbound flows from plants to DCs. The tradeoff among the fixed location and capacity acquisition costs, transportation cost and response time in two-echelon supply chain is complex, however, we make the following observations from the optimal solution:

- As response time cost increases relative to plant and DC location costs and transportation cost (i.e. $\theta \to \infty$), the model recommends using DCs with higher capacity level or in some cases more DCs are opened. The customer demand gets reallocated to decrease the average DC utilization, thereby improving the overall response time. More customers are assigned to their closest open DCs and the DCs are served by their closet open plants, thereby decreasing the inbound and outbound transportation cost. Note that demand splitting occurs in an attempt to balance the workload among open DCs.

- As the fixed location cost of DCs increases relative to the response time cost and transportation cost, the optimal solution recommends closing some of the DCs and using existing DCs but with higher capacity level (due to economies of scale). This consolidation of workloads at few DCs increases the average DC utilization, thereby increasing the overall response time. Most of the customers are assigned to their closest open DCs. It is worthwhile noting that no demand splitting was observed in thic case (due to the consolidation of workloads at few DCs).

3.5.2 Performance of the Lagrangean Heuristic

In Table 3.1, we compare the performance of the cutting plane algorithm and the Lagrangean heuristic and report the results for one problem set of the two-echelon ATO supply chain design problem (where $I = 100$, $J = 10$, $K = 3$, $M = 20$, and $N = 1$) for different values of the ratio of total plant capacities to total demand ($r = \sum_m P_m / \sum_i \lambda_i$). The results show that the Lagrangean heuristic outperforms the cutting plane method in terms of computational time. On average, the Lagrangean heuristic takes 421 sec whereas the cutting plane method takes 1344 sec. Furthermore, as we vary θ from 0.1, 1, 5, 10, 50, 100, 500, 1000, 2000, to 5000, the optimal solution demonstrates that substantial decrease in response time can be achieved (as a result of decrease in DC utilization due to workload reallocation and capacity acquisition) with a small increase in total cost associated with designing supply chains.

Table 3.2 shows the computational performance of the Lagrangean heuristic for M/G/1, M/M/1, and M/D/1 cases. The second subproblem pertaining to

Table 3.1: Comparison between the cutting plane method and the Lagrangean heuristic for ATO supply chain design for $I = 100$, $J = 10$, $K = 3$, $M = 20$, and $N = 1$.

Tight Capacities, $r = 3$

	Cutting Plane Method									Lagrangean Heuristic								
θ	FC	VC	RC	E(W)	$\bar{\rho}$	DC	CUT	ITR	CPU	FC	VC	RC	E(W)	$\bar{\rho}$	DC	LAG	GAP	CPU
	(%)	(%)	(%)						(s)	(%)	(%)	(%)				(%)	(%)	(s)
0.1	50	50	0	689.3	0.96	5	10	3	714	41	59	0	689.3	0.96	5	95.88	4.12	251
1	49	50	1	209.2	0.96	5	10	3	1029	41	59	1	218.2	0.96	5	96.78	3.22	268
5	48	50	2	169.3	0.95	5	10	3	1585	43	57	1	195.9	0.94	5	95.04	4.96	269
10	51	49	1	32.5	0.69	5	10	3	2180	42	56	2	51.84	0.76	5	95.81	4.19	271
50	51	47	2	14.51	0.6	5	5	2	1002	44	55	2	17.47	0.62	5	96.72	3.28	177
100	51	47	2	8.39	0.54	5	5	2	810	43	54	3	11.47	0.6	5	94.99	5.01	278
500	50	46	4	8.39	0.54	5	5	2	964	43	49	8	10.82	0.65	5	96.4	3.6	423
1000	49	44	7	6.21	0.44	5	5	2	1100	40	46	14	9.82	0.44	5	96.88	3.12	427
2000	46	41	13	5.99	0.44	5	5	2	1148	39	41	21	4.75	0.43	6	97.11	2.89	444
5000	38	32	30	3.95	0.32	7	28	5	1699	33	33	34	3.98	0.32	7	97.01	2.99	451

Moderate Capacities, $r = 5$

	Cutting Plane Method									Lagrangean Heuristic								
θ	FC	VC	RC	E(W)	$\bar{\rho}$	DC	CUT	ITR	CPU	FC	VC	RC	E(W)	$\bar{\rho}$	DC	LAG	GAP	CPU
	(%)	(%)	(%)						(s)	(%)	(%)	(%)				(%)	(%)	(s)
0.1	46	54	0	261.9	0.84	6	12	3	1500	39	61	0	332.3	0.85	4	98.11	1.89	205
1	46	54	0	98.12	0.83	6	6	2	964	39	61	0	122.7	0.84	4	96.83	3.17	407
5	41	58	1	34.16	0.83	4	4	2	1019	38	61	1	40.72	0.84	4	96.69	3.31	325
10	41	58	1	34.16	0.83	4	4	2	898	38	61	1	40.72	0.84	4	95.52	4.48	280
50	41	56	4	21.84	0.76	4	4	2	1467	39	58	3	27.56	0.79	4	97.27	2.73	322
100	43	55	2	6.81	0.56	4	4	2	1495	40	57	2	7.74	0.58	4	97.11	2.89	333
500	42	54	4	6.81	0.56	4	4	2	1122	37	53	10	7.74	0.58	4	96.88	3.12	365
1000	40	50	10	6.81	0.56	4	4	2	1370	34	48	18	7.74	0.58	4	94.01	5.99	291
2000	41	44	15	5.52	0.44	5	10	3	2524	40	38	22	6.63	0.47	6	95.6	4.4	589
5000	30	29	41	4.55	0.37	6	12	3	1981	30	29	40	6.53	0.44	6	97.67	2.33	232

Loose Capacities, $r = 10$

	Cutting Plane Method									Lagrangean Heuristic								
θ	FC	VC	RC	E(W)	$\bar{\rho}$	DC	CUT	ITR	CPU	FC	VC	RC	E(W)	$\bar{\rho}$	DC	LAG	GAP	CPU
	(%)	(%)	(%)						(s)	(%)	(%)	(%)				(%)	(%)	(s)
0.1	40	60	0	81.01	0.83	4	4	2	1103	40	60	0	81.66	0.83	4	96.57	3.43	546
1	40	60	0	62.75	0.83	4	4	2	955	39	60	0	63.85	0.83	4	96.65	3.35	435
5	40	60	1	40.27	0.83	4	4	2	906	39	60	1	42.23	0.83	4	98.91	1.09	487
10	39	59	1	40.27	0.83	4	4	2	875	39	60	1	41.78	0.83	4	94.41	5.59	327
50	40	58	2	11.71	0.67	4	4	2	979	40	58	2	12.97	0.67	4	96.19	3.81	354
100	40	56	4	11.71	0.67	4	4	2	1140	40	57	3	11.75	0.61	4	96.99	3.01	654
500	41	55	5	7.02	0.56	4	4	2	1080	38	52	11	8.96	0.56	4	97.27	2.73	765
1000	38	52	10	6.94	0.56	4	8	3	1765	34	47	19	7.84	0.56	4	98.1	1.9	642
2000	34	47	19	6.85	0.56	4	8	3	1884	41	38	21	7.54	0.37	6	95.59	4.41	822
5000	32	29	39	4.43	0.37	6	12	3	2148	31	29	39	4.43	0.37	6	96.37	3.63	984
min	30	29	0	3.95	0.32	4	4	2	714	30	29	0	3.98	0.32	4	94.01	1.09	177
max	51	60	41	689.3	0.96	7	28	5	2524	44	61	40	689.3	0.96	7	98.91	5.99	984
mean	43	50	7	63.38	0.66	5	7	2	1314	39	52	9	64	0.64	5	97.51	3.49	421

θ: Multiplier for the response time cost;
FC: Fixed cost; VC: Variable cost; RC: Response time cost (expressed as percentage of total cost);
$\mathbb{E}(W)$: Total expected waiting time;
$\bar{\rho}$: Average DC utilization;
DC: No. of DCs selected open;
CUT: Number of cuts generated;
ITR: Number of iterations;
CPU(s): CPU time in sec.
LAG: Lagrangean bound expressed as percentage of optimal solution;
GAP: 100×(Heuristic Solution-Optimal Solution)/Optimal Solution;
CPU: Total computational time (in sec).

the MTO echelon was solved to optimality using the cutting plane approach. In all these test problems, the heuristic is activated at the final iteration of the Lagrangean procedure. The Lagrangean bound (LAG) is expressed as the percentage of heuristic solution and the quality of the heuristic solution (GAP) is expressed as: 100 × (Heuristic Solution − LAG)/LAG. The table shows the computational time of the subproblems (SP), the master problem (MP) and the heuristic (H) expressed as a percentage of the total computational time (CPU) for various instances. From these results, it is evident that the proposed heuristic succeeds in finding feasible solutions that are within an average of 2.81%, 2.58%, and 2.99% of the Lagrangean bound in reasonable computational time: 427, 390, and 345 sec for M/G/1 case, M/M/1, and M/D/1 cases respectively. The total computational time can be as high as 1038 sec in some cases. In terms of the size of the test problems, the heuristic succeeds to solve problems with up to 35 plants, 20 DCs, and 150 customers, 5 products, 5 capacity levels and 5 semi-finished products within a maximum of 6% gap from the optimal solution. Table 3.2 shows that the solution of subproblem 2 accounts for most of the computational time, 89.08%, 89.50%, and 89.93% for M/G/1, M/M/1, and M/D/1 cases respectively. The master problem accounts for 9.87%, 9.36%, and 9.03%, whereas the heuristic accounts for 1.05%, 1.14%, and 1.04%, on average for M/G/1, M/M/1, and M/D/1 cases respectively.

3.6 Concluding Remark

In this chapter, we presented a model that captures the effect of response time reduction on the design of two-echelon ATO supply chain networks, that consists of plants and DCs serving a set of customers. Lagrangean relaxation was applied to decompose the problem by echelon - one for the MTS echelon and the other for the MTO echelon. While Lagrangean relaxation provides a lower bound, a heuristic is proposed that uses the solution of the subproblems to construct an overall feasible solution. Computational results reveal that the heuristic solution is on average within 6% from the optimal solution. We also used the models to demonstrate that substantial decrease in response time can be achieved with a small increase in total cost associated with designing supply chains.

Table 3.2: Computational performance of Lagrangean heuristic: ATO supply chain design

								M/G/1 Case(CV = 1.5)						M/M/1 Case						M/D/1 Case					
No.	I	J	K	L	M	N	θ	LAG(%)	GAP(%)	SP(%)	MP(%)	H(%)	CPU(s)	LAG(%)	GAP(%)	SP(%)	MP(%)	H(%)	CPU(s)	LAG(%)	GAP(%)	SP(%)	MP(%)	H(%)	CPU(s)
1	50	10	3	1	10	3	0.1	99.78	0.22	84.31	14.36	1.33	108	96.83	3.17	86.06	12.75	1.19	85	98.33	1.67	92.96	6.29	0.75	56
							1	95.84	4.16	84.55	13.51	1.94	111	99.59	0.41	85.21	12.96	1.83	85	94.58	5.42	91.34	8.06	0.60	62
							5	94.77	5.23	87.1	11.51	1.39	118	99.97	0.03	86.13	12.19	1.68	86	98.84	1.16	92.05	5.94	2.01	69
							10	98.33	1.67	83	14.98	2.02	121	95.77	4.23	93.14	5.53	1.33	88	96.94	3.06	88.43	10.90	0.67	71
							50	98.91	1.09	91.52	6.84	1.64	122	94.12	5.88	89.7	8.82	1.48	101	99.86	0.14	88.58	9.59	1.83	73
							100	97.20	2.80	88.78	11.13	0.09	127	95.40	4.6	91.73	7.01	1.26	104	98.65	1.35	90.91	7.93	1.16	75
2	50	20	3	1	10	3	0.1	97.07	2.93	87.85	11.17	0.98	191	99.66	0.34	85.99	12.41	1.6	176	98.23	1.77	93.59	5.44	0.97	131
							1	98.24	1.76	85.03	13.15	1.82	192	96.36	3.64	91.17	8.29	0.54	179	94.79	5.21	91.87	6.36	1.77	147
							5	99.32	0.68	85.22	13.04	1.74	199	99.07	0.93	92.56	6.14	1.3	181	95.92	4.08	94.09	5.39	0.52	148
							10	97.89	2.11	88.68	10.73	0.59	209	99.37	0.63	90.55	8.59	0.86	185	99.73	0.27	86.77	11.58	1.65	155
							50	94.06	5.94	94.66	5.17	0.17	211	99.02	0.98	87.18	12.46	0.36	186	97.71	2.29	91.02	6.97	2.01	158
							100	94.65	5.35	89.75	9.67	0.58	213	95.09	4.91	92.35	7.59	0.06	189	97.05	2.95	93.33	6.03	0.64	162
3	100	5	3	3	20	5	1	99.97	0.03	93.78	5.68	0.54	182	98.35	1.65	89.18	10.14	0.68	168	98.56	1.44	86.09	12.09	1.82	150
							5	96.58	3.42	85.45	14.39	0.16	185	99.75	0.25	92.9	6.7	0.4	169	95.33	4.67	88.18	11.27	0.55	159
							10	97.23	2.77	86.79	12.64	0.57	187	94.26	5.74	87.14	10.9	1.96	171	95.07	4.93	86.89	11.50	1.61	159
							50	97.41	2.59	93.22	5.83	0.95	190	96.71	3.29	92.66	5.43	1.91	176	99.89	0.11	86.5	12.72	0.78	161
							100	99.68	0.32	93.11	6.64	0.25	194	99.22	0.78	92.25	6.42	1.33	180	98.85	1.15	88.02	11.86	0.12	165
							200	95.64	4.36	93.61	5.06	1.33	196	99.55	0.45	88.63	10.35	1.02	175	94.97	5.03	85.38	12.94	1.68	165
4	100	10	3	3	20	5	1	94.47	5.53	91.73	8.09	0.18	238	94.01	5.99	87.35	11.95	0.7	217	95.40	4.60	91.29	8.25	0.46	197
							5	94.51	5.49	90.31	8.26	1.43	268	95.93	4.07	92.1	7.29	0.61	218	94.21	5.79	86.72	11.72	1.56	201
							10	98.83	1.17	90.63	9.29	0.08	277	98.77	1.23	87.36	11.18	1.46	222	97.99	2.01	92.51	5.88	1.61	202
							50	97.22	2.78	86.47	11.87	1.66	284	95.41	4.59	86.51	12.94	0.55	232	99.58	0.42	87.82	11.84	0.34	204
							100	97.40	2.60	91.17	7.41	1.42	296	98.45	1.55	90.21	9.1	0.69	234	96.62	3.38	85.71	12.99	1.30	210
							200	95.94	4.06	85.5	13.22	1.28	298	97.45	2.55	90.95	7.79	1.26	234	94.80	5.20	91.81	6.52	1.67	215
5	100	20	3	3	20	5	1	98.97	1.03	85.81	12.14	2.05	297	98.08	1.92	87.23	10.92	1.85	283	95.40	4.60	92.22	5.79	1.99	251
							5	97.61	2.39	92.9	5.91	1.19	298	94.37	5.63	92.15	7.57	0.28	285	98.65	1.35	93.11	5.43	1.46	268
							10	95.82	4.18	87.46	11.13	1.41	300	97.12	2.88	87.54	10.71	1.75	285	99.61	0.39	88.22	11.10	0.68	269
							50	95.55	4.45	91.68	7.47	0.85	309	97.03	2.97	89.47	9.11	1.42	286	94.92	5.08	86.04	12.13	1.83	271
							100	99.12	0.88	85.59	12.88	1.53	313	98.66	1.34	87.43	11.79	0.78	289	94.12	5.88	86.81	12.00	1.19	277
							200	95.46	4.54	89.1	9.64	1.26	348	96.44	3.56	92.22	7.55	0.23	293	96.12	3.88	91.16	8.60	0.24	278
6	150	5	5	5	35	5	1	98.67	1.33	89.13	9.93	0.94	551	99.94	0.06	86.61	12.26	1.13	486	95.08	4.92	92.21	6.98	0.81	423
							5	98.35	1.65	85.54	13.96	0.50	563	94.75	5.25	86.2	11.91	1.89	499	95.08	4.92	93.19	5.55	1.26	427
							10	94.73	5.27	93.29	6.17	0.54	576	96.05	3.95	90.43	9.19	0.38	513	98.69	1.31	87.81	10.86	1.33	444
							50	99.83	0.17	89.76	9.06	1.18	593	98.66	1.34	86.77	11.68	1.55	536	99.16	0.84	90.33	9.11	0.56	451
							100	99.01	0.99	85.81	13.66	0.53	599	96.90	3.1	92.06	7.38	0.56	544	97.22	2.78	91.59	7.71	0.70	484
							200	99.87	0.13	87.42	12.07	0.51	617	99.44	0.56	92.73	5.44	1.83	547	94.26	5.74	88.29	10.55	1.16	485
7	150	10	5	5	35	5	1	99.23	0.77	90.39	9.29	0.32	715	99.91	0.09	89.18	9.53	1.29	645	94.90	5.10	93.54	6.09	0.37	588
							5	97.23	2.77	91.5	7.63	0.87	717	99.00	1	89	9	2	690	96.47	3.53	90.24	9.10	0.66	598
							10	95.34	4.66	94.55	5.32	0.13	726	98.73	1.27	89.71	8.64	1.65	691	99.17	0.83	89.6	9.16	1.24	615
							50	97.13	2.87	90.74	7.41	1.85	741	98.15	1.85	91.6	7.61	0.79	697	95.27	4.73	88.82	10.47	0.71	624
							100	95.27	4.73	91.74	6.66	1.60	743	98.50	1.5	87.91	11.34	0.75	698	99.94	0.06	87.06	12.88	0.06	627
							200	98.12	1.88	91.44	7.15	1.41	747	96.97	3.03	91.45	7.14	1.41	706	97.41	2.59	92.49	6.96	0.55	632
8	150	20	5	5	35	5	1	98.31	1.69	91.9	6.98	1.12	981	95.47	4.53	87.58	11.25	1.17	922	95.07	4.93	94.1	5.13	0.77	765
							5	96.68	3.32	89.86	8.58	1.56	1000	99.81	0.19	91.22	7.21	1.57	935	95.90	4.10	85.95	12.40	1.65	800
							10	95.12	4.88	93.2	5.87	0.93	1004	96.02	3.98	87.96	10.64	1.4	945	98.35	1.65	92.89	6.46	0.65	810
							50	95.71	4.29	84.32	14.16	1.52	1007	99.50	0.5	91.95	7.83	0.22	955	94.35	5.65	86.9	13.03	0.07	871
							100	98.64	1.36	85.82	12.35	1.83	1019	94.38	5.62	89.96	9.18	0.86	972	99.94	0.06	90.32	9.60	0.08	890
							200	94.38	5.62	84.72	14.67	0.61	1038	94.09	5.91	88.84	9.49	1.67	976	99.72	0.28	92.09	6.17	1.74	904
							min	94.06	0.03	83.00	5.06	0.08	108	99.97	0.03	85.21	5.43	0.06	85	94.12	0.06	85.38	5.13	0.06	56
							max	99.97	5.94	94.66	14.98	2.05	1038	94.01	5.99	93.14	12.96	2.00	976	99.94	5.88	94.10	13.03	2.01	904
							mean	97.19	2.81	89.08	9.87	1.05	427	97.42	2.58	89.50	9.36	1.14	390	97.01	2.99	89.93	9.03	1.04	345

I: No. of customers; J: No. of potential DCs; K: No. of capacity levels at each DC; L: No. of (finished) products, M: No. of plants, N: No. of semi-finished products;
θ: Multiplier for the response time costs;
LAG: Lagrangean bound expressed as percentage of optimal solution;
GAP: 100×(Heuristic Solution-Optimal Solution)/Optimal Solution;
SP: Computational times of the subproblems; MP: Computational times of the master problems; H: Computational times of the heuristics;
CPU: Total computational time in seconds; (Note that SP, MP, and H are expressed as percentages of the total computational time (CPU).

Chapter 4

Service-Level Differentiation in MTO Supply Chain Design for Segmented Markets[‡]

4.1 Introduction

Service-level differentiation is an emerging strategy adopted by manufacturing and service firms operating in an MTO environment to manage customer segments with different profitability and different service quality requirements. Firms segment customers into different classes to which they offer the same product or service but with different levels of service quality so as to maximize profits. Furthermore, in supply chains that support after-sales services, the delivery of differentiated levels of service to disparate classes of customers is an increasingly important requirement in today's customercentric environment. In make-to-stock systems, service-level differentiation can be an effective way of utilizing inventory investment as they provide higher service levels for the more critical parts at the expense of accepting lower service levels for parts with less impact [51].

[‡] A version of the materials presented in this chapter will be submitted for publication: N. K. Vidyarthi, S. Elhedhli, and E. M. Jewkes, (2009) Demand allocation and capacity decisions in make-to-order supply chain design for segmented markets with service-level differentiated customers [146].

In this chapter, our primary objective is to present a model that seeks to simultaneously determine the location and the service capacity of DCs and allocate demand arising from various customers to DCs in an MTO supply chain that serves multiple classes of customer with different service-level requirements. Specifically, we study the following issues: How will an MTO supply chain be designed for service level differentiated customers? How will the demand be allocated and service capacity be segmented for different customer classes? In MTO systems, where no finished product inventory is held in the system, the service quality is often specified as a function of order-to-delivery lead time or response time. It is worthwhile noting that in the literature, the service level is often specified in terms of *expected waiting time* of customers mainly for tractability reasons (see Silva and Serra [126] and references therein). Placing bounds on *expected waiting time* reduces congestion in the system, however it does not guarantee that the actual waiting time of the customer is within a pre-specified limit. If the distribution of the service-time has a long-tail, then the system can frequently generate waiting times that are way above the average waiting times [1]. In such cases, constraints involving expected waiting time measures can lead to suboptimal solutions. Therefore, in this research, we express service level of a customer class in terms of the fraction of demand served within a pre-specified waiting time (sojourn time) for that class. To the best of our knowledge, none of the models proposed in the literature has considered the design of such service/supply chain systems that provides a *waiting-time based* service-level guarantee on the customer order in the presence of multiple customer classes whereas research (Batta, 1989) has shown that the optimal location of facilities under multiple customer classes is usually different from that obtained by grouping demand arrivals from all priorities into a single category.

We present an MTO supply chain design model that seeks to simultaneously determine the location and the capacity of distribution centers (DCs) and allocate stochastic customer demand arising from multiple customer classes to DCs by minimizing the fixed cost of opening DCs and equipping them with sufficient assembly capacity and the variable cost of serving customers subject to service level constraints on response time for each of these customer classes. Note that customer classes vary in their demand rates and service level requirements.

The rest of the chapter is organized as follows. In Section 4.2, we review related literature. We present the problem description, the mathematical formulation, and the linearization of the model in Section 4.3. Section 4.4 describes the procedure for estimation of the service level function and its subgradients. In Section 4.5, we present a solution procedure based on cutting plane algorithm. Computational results and insights are reported in Section 4.6. Finally, Section 4.7 provides the concluding remark.

4.2 Related Literature

Multiple customer classes have received increased attention in the operations research/ operations management literature during the last several years. Supply chains are often characterized by multiple customer segments differentiated by their service-level requirements, e.g. service parts logistics systems, airlines, and hotels. The practice of inventory rationing, i.e. issuing stock to some customers, while refusing (lost sales) or delaying demand fulfillment (backordering) for other customers setting, as a way to cope with multiple customer segments having different expectations, has been studied extensively in the literature. This includes research on inventory rationing in periodic review systems (see Frank et al. [65] and references therein), for continuous-review settings (see Deshpande et al. [52] and its citations therein), the case of make-to-stock systems (see Ha [70, 71, 72] and bibliographies there), service parts logistics systems [150], and production-inventory in general [55]. All of these models consider situations, where inventory of finished products are held in stock at a facility, and the customer demand (if fulfilled) is met from the stock. The research presented in this thesis differs from the existing literature on multiple customer classes in that we deal with the design of MTO system where the processing/assembly capacity multiple facility locations is *rationed* among different customer classes, so as to satisfy the demand with prespecified response-time-based service-level requirements that may differ amongst the customer classes.

Another body of literature that is relevant to our work is *stochastic location* *models with immobile server and priority class.* One of the early papers in this area is Batta et al. [19]. Batta et al. [19] point that queuing disciplines frequently used in decision models (such as first-come first-served, last-come-first-served, and service-in-random-order) are clearly inappropriate in many contexts (e.g. urban emergency services, police patrols). They present a formulation and solution techniques for a single server priority queueing location model (PQL) that allow calls to be selected from an arbitrary number of priority classes. Furthermore, they show that the optimal K-PQL model prescribes location that is usually different from that obtained by grouping arrivals from all priorities into a single category and using the single queue length model. Batta [18] considers the problem of locating a single server on a network operating as an M/G/1 queue, in which queued calls are serviced by a class of queuing disciplines that depend solely on expected service time information. The model is analyzed as an M/G/1 non-preemptive priority queuing model, with location-dependent priorities. More recently, Silva and Serra [126] present model and solution algorithms for priority queue covering location problem (PQCLP) which seeks to locate (emergency) service facilities when the arrivals have different priorities. Their model maximizes the population covered while ensuring that constraints on the average waiting time for each customer class are met.

4.3 Model Formulation

We consider the problem of designing an MTO supply chain (Figure 4.1) consisting of J potential DC locations (production or assembly facilities) that serves I customer locations with the demand for a single product. These customers belong to one of N *priority classes* (class 1 has the highest priority and class N being the lowest) and these classes may vary in their demand rates and service level requirements. Demand from each customer class n ($n = 1, ..., N(i)$) from location i ($i = 1, ..., I$) occurs one unit at a time according to an independent Poisson process with mean arrival rate λ_i^n. Once the demand is realized at the customers' end, the order is placed to the DC. The DC acts as an processing facility that consists of a server with infinite buffer to accommodate customer orders waiting for service. The DC begins the processing of the final product after the receipt of an order. Orders from different customers are processed at the DC in the order of priority classes, but on a first-come-first-serve (FCFS) basis within each priority class. Because the system operates on a MTO basis, no finished product inventory is held at any point in the system, and hence the customer orders cannot be met immediately.

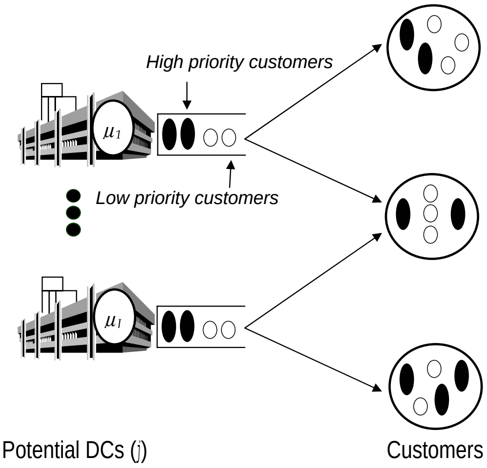


Figure 4.1: Schematic MTO supply chain with multiple customer classes

Let c_{ij} be the cost of fulfilling an order (unit transportation and/or production cost) of customer i from DC j ($j = 1, ..., J$). Processing times at the DCs are assumed to be exponentially distributed with the mean $1/\mu_{jk}$, if DC j is equipped with a service capacity level k ($k = 1, ...K$). We model the flexible capacity of each DCs as a single server with a choice of $k = 1, ...K$ discrete capacity levels (with the corresponding service rate μ_{jk}). We assume that the processing time at a DC is independent of the customer class, however the model can be easily extended to deal with that. Let f_{jk} be the fixed cost (amortized over the planning period) associated with the use of DC j that is equipped with capacity level k. If x_{ij}^n denote the long-run fraction of demand of customer i of class n served by DC j ($0 \leq x_{ij}^n \leq 1$), then the total demand served by DC j is also a random variable that follows a Poisson process with mean $\lambda_j = \sum_{i=1}^{I} \sum_{n}^{N(i)} \lambda_i^n x_{ij}^n$. Furthermore, we introduce a binary decision variable y_{jk} that takes the value 1, if DC j equipped with capacity level k is used and 0 otherwise. To ensure the overall stability of the system, we assume that for every DC j, the total demand served is less that its service capacity: $\sum_{i=1}^{I} \sum_{n=1}^{N} \lambda_i^n x_{ij}^n \leq \sum_{k=1}^{K} \mu_{jk} y_{jk}$. To ensure that the demand for each product is met, we require that $\sum_{j=1}^{J} x_{ij}^n = 1$. Hence, each DC j can be modelled as an $M/M/1$ priority queue with aggregate arrival rate λ_j and service rate μ_j and the system can be viewed as a network of spatially distributed $M/M/1$ priority queues. Our objective is to simultaneously determine the allocation of the multiclass customer demand to the DCs, $\mathbf{x}^*$ and the location and capacity level of the DCs $\mathbf{y}^*$ so as to minimize the fixed cost of using DCs equipped with sufficient processing capacity and the variable cost of fulfilling customer demands subject to meeting the specified service level requirements.

In an MTO system, where customer orders cannot be met immediately, it is reasonable to specify the service level requirement as a function of waiting time of the orders. Service-level constraints are often specified as bounds on the average waiting time of a customer in the system [126]. Note that specifying bounds on the average waiting time of customers in the system *does not guarantee* that the actual waiting times of customers are less than the average waiting time specially when the waiting-times has long-tail distribution. For example, consider an MTO supply chain configuration for a single class of customers, M/M/1 case, $\theta = 0.1$ from Table 2.1. The optimal configuration prescribes opening 4 DCs, each with a capacity of $\mu_1 = 13700$, $\mu_2 = 13700$, $\mu_3 = 13700$, and $\mu_4 = 6850$. The total demand allocated to these DCs are $\lambda_1 = 13289$, $\lambda_2 = 13563$, $\lambda_3 = 12056$, and $\lambda_4 = 6781.5$. The expected waiting time at the DCs are $W_1 = 33.93$, $W_2 = 98.34$, $W_3 = 7.24$, and $W_4 = 74.55$. Let us place a bound of 100 units on the waiting time ($\tau = 100$). Hence, the probabilities that the actual waiting time of a customer at the DCs is less than 100 units ($Pr(W_j \leq 100)$ can be computed as follows: $Pr_1 = 1 - \exp^{-(\mu_1 - \lambda_1)/100} = 0.98$, $Pr_2 = 0.75$, $Pr_3 = 0.99$, $Pr_4 = 0.49$. From these probabilities, it is clear that a service-level of $Pr_2 = 0.75$ and $Pr_4 = 0.49$ is unacceptable, although the average waiting time at every DCs is within the upper bound (of 100 units). Hence, we specify the service level requirements as the fraction of demand served within a specified response (sojourn) time. This can expressed as the probability that a customer order from a priority class spends more than τ time units does not exceed α for some finite τ and $\alpha \in (0, 1)$. For a given demand allocation $\mathbf{x}$ and capacity level $\mathbf{y}$, let the arrival rate and service rate at a DC j be denoted by $\lambda_j^n(\mathbf{x})$ and $\mu_j(\mathbf{y})$. If we let $W_j^n(\lambda_j^n, \mu_j)$ denote the total time spent in the system (waiting in queue + service time) by an order of class n at DC j, and τ_j^n is the quoted response time, then the service level constraint can be expressed as follows:

$$S_j^n(\mathbf{x}, \mathbf{y}, \tau^n) = Pr\{W_j^n(\lambda_j^n, \mu_j) \leq \tau^n\} \geq \alpha^n \qquad \forall j, n$$

where $\alpha^n \in [0, 1]$ is the specified service level for customer class n at DC j, $\lambda_j^n(\mathbf{x}) = \sum_{i=1}^{I} \lambda_i^n x_{ij}^n$ and $\mu_j(\mathbf{y}) = \sum_{k=1}^{K} \mu_{jk} y_{jk}$. In other words, $S_j^n(.)$ are the cumulative distribution functions (CDFs) of $W_j^n(.)$.

With these notations, the MTO supply chain design problem for multiple customer classes with service-level constraints can be formulated as follows:

$$
\begin{aligned}
[MC] : \min z(\mathbf{x}, \mathbf{y}) = & \sum_{j=1}^{J} \sum_{k=1}^{K} f_{jk} y_{jk} + \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{n=1}^{N} c_{ij} \lambda_i^n x_{ij}^n && && (4.1) \\
\text{s.t.} \quad & \sum_{i=1}^{I} \sum_{n=1}^{N} \lambda_i^n x_{ij}^n \leq \sum_{k=1}^{K} \mu_{jk} y_{jk} && \forall j && (4.2) \\
& \sum_{k=1}^{K} y_{jk} \leq 1 && \forall j && (4.3) \\
& \sum_{j=1}^{J} x_{ij}^n = 1 && \forall i, n && (4.4) \\
& Pr\{W_j^n(\mathbf{x}, \mathbf{y}) \leq \tau^n\} \geq \alpha^n && \forall j, n && (4.5) \\
& 0 \leq x_{ij}^n \leq 1, \quad y_{jk} \in \{0, 1\} && \forall i, j, k, n && (4.6)
\end{aligned}
$$

The objective function (4.1) minimizes the sum of fixed cost of locating DCs and acquiring sufficient capacity and the variable cost of assembly and shipment of products from DCs to the customers. Constraints (4.2) are the stability condition for the queuing system, which models the DCs. Constraint set (4.3) ensures that at most one capacity level is selected at a DC, whereas constraint set (4.4) ensures that the demand for each customer is met. Constraint set (4.5) are the service level requirements for the various classes of customers. Constraint set (4.6) are the nonnegativity and binary constraints.

The underlying model is difficult to solve due to the lack of closed form expression for service-level constraint (4.5) for multiple customer classes . Therefore we will use simulation where the service-level functions $S_j^n(\mathbf{x}, \mathbf{y}, \tau^n)$ are estimated by corresponding sample averages $\widehat{S}_j^n(\mathbf{x}, \mathbf{y}, \tau^n, m)$, where m is the sample size used. Furthermore, the above model seems to be linear except for the service-level function $\widehat{S}_j^n(\mathbf{x}, \mathbf{y}, \tau^n, m)$. Our initial testing shows that the components of $\widehat{S}_j^n(\mathbf{x}, \mathbf{y}, \tau^n, m)$ are concave. Intuitively, one would expect that the service level increases with decreasing marginal returns as the service rate increases. Furthermore, it should decrease with increasing marginal returns as the arrival rate increases. Our initial simulation results show that this is a reasonable assumption. Chen and Henderson [39] have also shown that in an M/M/s queue, the distribution of steady state waiting time of customers evaluated at any fixed value, is concave and decreasing function of arrival rate. If this concavity assumption holds, then we can approximate the service level function with piecewise linear concave function.

For the sake of clarity, we will limit our analysis and discussion to two customer classes ($n = h$ and l: h being the high priority class and l being the low priority class) in the remainder of the section. However, without loss of generality the model remains valid for N classes. Furthermore, it is worthwhile noting that we assume *preemptive priority* queue because of the existence of closed form solution for the tail of response time distribution for the high priority customers. For the low priority customers, we rely on simulation or matrix analytic method for the estimation of the response time distribution. For the *non-preemptive priority case*, one has to rely on simulation or matrix analytic method for the estimation of the response time distribution for both classes of customers (due to lack of closed-form expressions).

4.3.1 Linearization

The tail of the response time distribution $S_j^h(.)$ for *high priority customers* in a *preemptive priority queue* is known to be exponential [154] and is given by:

$$S_j^h(.) = Pr(W_j^h \leq \tau^h) = 1 - e^{-(\mu_j - \lambda_j^h)\tau_j^h}$$

Using this, the service level constraint (4.5) for the high priority customer can be expressed as a linear constraint:

$$\sum_{k=1}^{K} \mu_{jk} y_{jk} - \sum_{i=1}^{I} \lambda_i^h x_{ij}^h \geq \frac{-\ln(1-\alpha^h)}{w_j^h} \qquad \forall j \tag{4.7}$$

If the concavity assumption holds, then the service level for *low priority customers* $S_j^l(.)$ can be approximated by a set of supporting hyperplanes that are tangent to $S_j^l(.)$ at various points $(\lambda_j^h, \lambda_j^l, \mu_j^h), \forall q \in Q$, that is

$$S_j^l(.) = \min_{q \in Q} \left\{ S_j^{lq}(.) + (\lambda_j^l - \lambda_j^{lq}) \left(\frac{\partial S_j^{lq}(.)}{\partial \lambda_j^l} \right) + (\lambda_j^h - \lambda_j^{hq}) \left(\frac{\partial S_j^q(.)}{\partial \lambda_j^h} \right) + (\mu_j - \mu_j^q) \left(\frac{\partial S_j^{lq}(.)}{\partial \mu_j} \right) \right\} \quad \forall j$$

This can be written as

$$S_j^l(.) \leq \quad S_j^{lq}(.) + (\lambda_j^l - \lambda_j^{lq}) \left(\frac{\partial S_j^{lq}(.)}{\partial \lambda_j^l} \right) + (\lambda_j^h - \lambda_j^{hq}) \left(\frac{\partial S_j^q(.)}{\partial \lambda_j^h} \right) + (\mu_j - \mu_j^q) \left(\frac{\partial S_j^{lq}(.)}{\partial \mu_j} \right) \quad \forall j, q \in Q$$

where $\frac{\partial S_j^{lq}(.)}{\partial \lambda_j^h}$, $\frac{\partial S_j^{lq}(.)}{\partial \lambda_j^l}$ and $\frac{\partial S_j^{lq}(.)}{\partial \mu_j}$ are the subgradients of $S_j^{lq}(.)$ at points $(\lambda_j^h,\ \lambda_j^l,\ \mu_j)$.

This implies that the service level constraint for the low priority customer can be expressed as a set of linear constraints as follows:

$$\frac{\partial S_j^{lq}(.)}{\partial \lambda_j^l} \sum_{i=1}^{I} \lambda_i^l x_{ij}^l + \frac{\partial S_j^{lq}(.)}{\partial \lambda_j^h} \sum_{i=1}^{I} \lambda_i^h x_{ij}^h + \frac{\partial S_j^{lq}(.)}{\partial \mu_j} \sum_{k=1}^{K} \mu_{jk} y_{jk} \geq$$

$$- S_j^{lq}(.) + \lambda_j^{lq} \frac{\partial S_j^{lq}(.)}{\partial \lambda_j^l} + \lambda_j^{hq} \frac{\partial S_j^{lq}(.)}{\partial \lambda_j^h} + \mu_j^q \frac{\partial S_j^{lq}(.)}{\partial \mu_j} + \alpha^n \quad \forall j, q \in Q$$

The resulting linear MIP model $[MC_L]$ is as follows:

$$\min_{x,y} \quad \sum_{j=1}^{J}\sum_{k=1}^{K} f_{jk}y_{jk} + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n\in(l,h)} c_{ij}\lambda_i^n x_{ij}^n \tag{4.8}$$

$$\text{s.t.} \quad \sum_{i=1}^{I}\sum_{n\in(l,h)} \lambda_i^n x_{ij}^n \le \sum_{k=1}^{K}\mu_{jk}y_{jk} \qquad \forall j \tag{4.9}$$

$$\sum_{k=1}^{K} y_{jk} \le 1 \qquad \forall j \tag{4.10}$$

$$\sum_{j=1}^{J} x_{ij}^n = 1 \qquad \forall i, n \tag{4.11}$$

$$\sum_{k=1}^{K}\mu_{jk}y_{jk} - \sum_{i=1}^{I}\lambda_i^h x_{ij}^h \ge \frac{-\ln(1-\alpha_j^h)}{w_j^h} \qquad \forall j \tag{4.12}$$

$$\begin{aligned} &\frac{\partial S_j^{lq}(.)}{\partial \lambda_j^l}\sum_{i=1}^{I}\lambda_i^l x_{ij}^l + \frac{\partial S_j^{lq}(.)}{\partial \lambda_j^h}\sum_{i=1}^{I}\lambda_i^h x_{ij}^h + \frac{\partial S_j^{lq}(.)}{\partial \mu_j}\sum_{k=1}^{K}\mu_{jk}y_{jk} \ge \\ &\qquad - S_j^{lq}(.) + \lambda_j^{lq}\frac{\partial S_j^{lq}(.)}{\partial \lambda_j^l} + \lambda_j^{hq}\frac{\partial S_j^{lq}(.)}{\partial \lambda_j^h} + \mu_j^q\frac{\partial S_j^{lq}(.)}{\partial \mu_j} + \alpha^n \qquad \forall j, q \in Q \end{aligned} \tag{4.13}$$

$$0 \le x_{ij}^n \le 1, \quad y_{jk} \in \{0,1\} \qquad \forall i, j, k, n \tag{4.14}$$

In the next section, we describe the procedure for estimating the service level $S_j^l(.)$ and its subgradients.

4.4 Estimation of Service Level Function and its Subgradients

In this section, we describe two procedures for estimating the service level $S_j^l(.)$ and its subgradients; the first is based on discrete-event simulation, and the second uses matrix analytic methods [83, 99]. The matrix-analytic methods can provide *near-exact* estimates of service level function in some cases.

4.4.1 Discrete-Event Simulation

Service Level Function: We use discrete-event simulation to estimate the value of service level function $S_j^l(.)$ at a given point $(\lambda_j^h,\ \lambda_j^l,\ \mu_j)$. Suppose we run a simulation with sample size m, where we independently generate the realizations of aggregate arrival times $(1/\lambda_j^h,\ 1/\lambda_j^l)$ and service times $(1/\mu_j)$ from their respective distributions. Let ξ denote the set of all the random arrival and service times and let ξ^1,...,ξ^m denote the independent realizations of ξ. Let $K_j(\xi^d)$ denote the total number of orders from lower class customers and let $s_j(\xi^d, x, y)$ denote the number of orders from lower class customers whose response time is within a prespecified time limit τ^l, corresponding to the arrival and service rates at the DC j, based on the demand allocation x and capacity level y acquired. Then an estimate of the service level function, $S_j^l(.)$, measured as the fraction of customer orders receiving adequate service in the long run is given by

$$\widehat{S}_j^l(\lambda_j^h, \lambda_j^l, \mu_j, \tau^l) = \frac{\mathbb{E}[s_j]}{\mathbb{E}[K_j]} = \frac{\frac{1}{m}\sum_{d=1}^m s_j(\xi^d, x, y)}{\frac{1}{m}\sum_{d=1}^m K_j(\xi^d)} \qquad \forall j$$

To compute the function, we use simulation with common random numbers, i.e. make sure that the same random numbers are used for all values of demand allocation x and capacity level y acquired [85]. The convergence of the estimate $\widehat{S}_j^l(., m)$ to the actual value of $S_j^l(.)$ as $m \to \infty$ with probability one, can be proved using techniques very similar to Atlason [13] and Cezik and Ecuyer [38].

Subgradients of Service Level Function: Some of the methods for gradient estimation are finite difference method, perturbation analysis, likelihood ratio methods, frequency domain methods, and simultaneous perturbation method. In this chapter, we use finite difference method as this is the most straightforward and intuitive method for estimating subgradients, when an expression for the function is unknown. Furthermore, the finite difference method has been shown to provide better estimates of gradients despite the fact that it requires more simulation runs compared to other methods such as simultaneous perturbation and likelihood ratio methods [11, 13, 14]. Gradient estimation through finite difference method can be obtained using forward differences, backward differences, or central differences. We choose to use central differences as they usually provide an estimate that has less bias than the forward or backward differences [11, 94].

In order to estimate the subgradients of a function (i.e. partial derivatives with respect to a continuous variable) using central finite differences, the function is evaluated at two different points. Then an estimate of the partial derivative at a particular value can be found by linear interpolation. If the variable is integer, then the smallest difference between the two points is one. In our case, the arrival rate λ_j is a continuous variable as $0 \leq x_{ij} \leq 1$, and service rate μ_j is a discrete variable as $y_{jk} \in \{0,1\}$. If $\frac{\partial S_j^l(\lambda_j^h,\lambda_j^l,\mu_j)}{\partial \lambda_j^h}$, $\frac{\partial S_j^l(\lambda_j^h,\lambda_j^l,\mu_j)}{\partial \lambda_j^l}$, and $\frac{\partial S_j^l(\lambda_j^h,\lambda_j^l,\mu_j)}{\partial \mu}$ denote the subgradient of $S_j^l\left(\lambda_j^h,\lambda_j^l,\mu_j\right)$, then the central finite difference estimate are obtained as follows:

$$\begin{aligned}
\frac{\partial S_j^l(\lambda_j^h,\lambda_j^l,\mu_j)}{\partial \lambda_j^h} &\simeq \frac{\widehat{S}_j^l\left(\lambda_j^h+d\lambda_j^h,\lambda_j^l,\mu_j\right)-\widehat{S}_j^l\left(\lambda_j^h-d\lambda_j^h,\lambda_j^l,\mu_j\right)}{2d\lambda_j^h} && \forall j\\
\frac{\partial S_j^l(\lambda_j^h,\lambda_j^l,\mu_j)}{\partial \lambda_j^l} &\simeq \frac{\widehat{S}_j^l\left(\lambda_j^h,\lambda_j^l+d\lambda_j^l,\mu_j\right)-\widehat{S}_j^l\left(\lambda_j^h,\lambda_j^l-d\lambda_j^l,\mu_j\right)}{2d\lambda_j^l} && \forall j\\
\frac{\partial S_j^l(\lambda_j^h,\lambda_j^l,\mu_j)}{\partial \mu} &\simeq \frac{\widehat{S}_j^l\left(\lambda_j^h,\lambda_j^l,\mu_j+d\mu_j\right)-\widehat{S}_j^l\left(\lambda_j^h,\lambda_j^l,\mu_j-d\mu_j\right)}{2d\mu} && \forall j
\end{aligned}$$

where $d\lambda_j^h$, $d\lambda_j^l$ and $d\mu_j$ (referred as step size) are the incremental change in arrival rate of high priority, arrival rate of low priority and service rate of customers at DC j respectively. Note that the symbols $\widehat{S}_j^l(.)$ denote the estimates of $S_j^l(.)$ obtained from simulation at their corresponding parameter values. It is clear that we would conduct six simulation runs to obtain these three estimates of subgradients at a point $(\lambda_j^h,\lambda_j^l,\mu_j)$ for every DC j that is selected open ($y_{jk}=1$). These estimates of subgradients are used to generate the constraints of the form (4.13).

4.4.2 Matrix Analytic Method

Alternatively, one can use matrix analytic methods in some cases to estimate the service level function for the low priority customers. Details regarding this method can be found in Latouche and Ramaswami [83] and Neuts [99].

Let us determine the joint distribution of queue lengths. For that, let the number of high and low priority customers in the system (including the one in the service) be denoted by N_h and N_l, respectively. We assume that $N_l \geq 0$ (infinite low priority class buffer size) whereas $0 \leq n_h \leq M$ (the buffer size of the high priority customers in the system be M). No other state variables are required to model the system since the service is exponential and it is not necessary to keep track of which type of customer the server is attending to. As long as there is at least one high priority customer present in the system, the system must be busy attending to high priority queue. Therefore, the vector $\mathbf{N} = (N_l, N_h)$ represents states of a continuous-time Markov chain with state space $\{\mathbf{n} = (n_l, n_h) | n_l \geq 0, \quad 0 \leq n_h \leq M\}$. In the Markov process $\{\mathbf{N}\}$, a transition can occur only if a customer of either class arrives or a customer of either class is served. For example, with the arrival of a high priority customer with rate λ_h, the system transits from state $\mathbf{n}$ to $\mathbf{n}'$, where $\mathbf{n}' = \{(n_l, n_h + 1\}$ and with the arrival of a low priority customer with rate λ_l, the system transits from state $\mathbf{n}$ to $\mathbf{n}'' = \{(n_l + 1, n_h)\}$. Similarly, with the service of a high priority customer with rate μ_h, the system transits from state $\mathbf{n} = \{(n_l, n_h) | n_l \geq 0, n_h > 0\}$ to $\dot{\mathbf{n}} = \{(n_l, n_h - 1)\}$ and with the service of a low priority customer with rate μ_l, the system transits from state $\mathbf{n} = \{(n_l, n_h) | n_l \geq 0, n_h = 0\}$ to $\ddot{\mathbf{n}}$, where $\ddot{\mathbf{n}} = \{(n_l - 1, n_h)\}$. We order the states of the system lexicographically, i.e. $(0,0),(0,1),(0,2),\ldots,(0,M)$; $(1,0),(1,1),(1,2),\ldots,(1,M)$; $\ldots$; $(i,0),(i,1),(i,2),\ldots,(i,M)$, and define $\pi_{(i,s)}$ to be the stationary probability of the state (i,s). First of all, let us determine the steady-state joint distribution of queue lengths, that can be represented by $\pi \equiv (\underline{\pi}_0, \underline{\pi}_1, \underline{\pi}_2, \underline{\pi}_3, ...)$, where $\underline{\pi}_i \equiv (\pi_{(i,0)}, \pi_{(i,1)}, \pi_{(i,2)}, \pi_{(i,3)}, ..., \pi_{(i,M)})$. With n_l serving as the level and n_h as the sub-level, the infinitesimal generator of the chain $\mathbf{N}$ for $n_l = 0,1,2$ and $n_h = 0,1,...,M$ is given by:

$$Q = \left(\begin{array}{c|cccc|cccc|cccc}
 & (0,0) & (0,1) & (0,...) & (0,M) & (1,0) & (1,1) & (1,...) & (1,M) & (2,0) & (2,1) & (2,...) & (2,M) \\
(0,0) & -\delta_1 & \lambda_h & & & \lambda_l & & & & & & & \\
(0,1) & \mu & -\delta_2 & \lambda_h & & & \lambda_l & & & & & & \\
(0,...) & & \mu & -\delta_2 & \lambda_h & & & \lambda_l & & & & & \\
(0,M) & & & \mu & -\delta_3 & & & & \lambda_l & & & & \\
\hline
(1,0) & \mu & & & & -\delta_2 & \lambda_h & & & \lambda_l & & & \\
(1,1) & & & & & \mu & -\delta_2 & \lambda_h & & & \lambda_l & & \\
(1,...) & & & & & & \mu & -\delta_2 & \lambda_h & & & \lambda_l & \\
(1,M) & & & & & & & \mu & -\delta_3 & & & & \lambda_l \\
\hline
(2,0) & & & & & \mu & & & & -\delta_2 & \lambda_h & & \\
(2,1) & & & & & & & & & \mu & -\delta_2 & \lambda_h & \\
(2,...) & & & & & & & & & & \mu & -\delta_2 & \lambda_h \\
(2,M) & & & & & & & & & & & \mu & -\delta_3
\end{array}\right)$$

where $\delta_1 = \lambda_h + \lambda_l$, $\delta_2 = \lambda_h + \lambda_l + \mu$, and $\delta_3 = \mu + \lambda_l$. The entries of the generator matrix can be grouped into blocks to form a block-tridiagonal matrix as follows:

$$Q = \begin{pmatrix}
B_0 & A_1 & & & & \\
A_{-1} & A_0 & A_1 & & & \\
 & A_{-1} & A_0 & A_1 & & \\
 & & A_{-1} & A_0 & A_1 & \\
 & & & \ddots & \ddots & \ddots
\end{pmatrix}$$

where B_0, A_1, A_0, A_{-1} are square matrices of order $M+1$ defined as:

$$B_0 = \begin{pmatrix}
-\delta_1 & \lambda_h & & & \\
\mu & -\delta_2 & \lambda_h & & \\
 & \ddots & \ddots & \ddots & \\
 & & \mu & -\delta_2 & \lambda_h \\
 & & & \mu & -\delta_3
\end{pmatrix}; \quad A_1 = \begin{pmatrix}
\lambda_l & & & & \\
 & \lambda_l & & & \\
 & & \ddots & & \\
 & & & \ddots & \\
 & & & & \lambda_l
\end{pmatrix}; \quad A_{-1} = \begin{pmatrix}
\mu & & & & \\
 & 0 & & & \\
 & & \ddots & & \\
 & & & \ddots & \\
 & & & & 0
\end{pmatrix}$$

and $A_0 = B_0 - A_{-1}$.

The matrix B_0 contains all transitions when no low priority customers are present in the system and the server is devote to serving high priority customers. A_1 contains all transitions that represents arrivals of low priority customers, whereas A_{-1} contains transitions corresponding to the service of low priority customer. Since n_l can only change by ±1, the only non-zero matrices are A_1, A_0, and A_{-1}. As a result, the system under consideration is a continuous-time *quasi-birth-and-death* (QBD) process. Thus, using $\pi Q = \mathbf{0}$, we have the steady-state balance equation in matrix form:

$$\pi_{\mathbf{0}} B_0 + \pi_{\mathbf{1}} A_{-1} = \mathbf{0}$$

$$\pi_{\mathbf{j-1}} A_1 + \pi_{\mathbf{j}} A_0 + \pi_{\mathbf{j+1}} A_{-1} = \mathbf{0} \qquad \forall j \geq 1$$

which can be written in the recursive form, i.e.

$$\pi_{\mathbf{j}} = \pi_{\mathbf{j-1}} R \qquad \forall j \geq 1$$

where the rate matrix R is the minimal non-negative solution to the quadratic equation:

$$A_1 + RA_0 + R^2 A_{-1} = \mathbf{0}$$

The steady-state probabilities $\pi_{\mathbf{0}}$ are determined from:

$$\pi_{\mathbf{0}}(A_1 + RA_{-1}) = \mathbf{0}$$

subject to the normalization equation:

$$\sum_{k=0}^{\infty} \pi_{\mathbf{k}} \mathbf{e} = \pi_{\mathbf{0}} (I - R)^{-1} \mathbf{e} = 1$$

where I denotes the identity matrix and $\mathbf{e}$ is a column vector of ones of size $M + 1$. These steady state probabilities will be used in estimating the service-level for low priority customers.

This matrix analytic procedure is very efficient for obtaining the near-exact performance measures through judicious choice of the number of states. Note that the computational implementation of the procedure requires that the number of states in the QBD process be finite. We begin by treating the queue length of the high and low priority customers to be of finite size but of sufficiently large size that the estimates of desired performance measures are quite accurate. However, the computational effort grows rapidly with the number of states and customer classes, making it necessary to rely on simulation.

4.4.2.1 Estimation of Service-Level for Low Priority Customers

We derive the distribution of response time of low priority customers. The response time of a low priority customer W_j^l is the time between its arrival to the system till it completes its service (i.e. waiting time in queue plus the time in service). We assume that the low priority customer may be *preempted* by one or more of the high priority customers for service. However, the method can be extended to deal with the non-preemptive priority case. In general, it is difficult to characterize the distribution of the service-level S_j^l in such systems. However, Ramaswami and Lucantoni [109] present an efficient algorithm for the derivation of complementary distribution of stationary waiting times in phase-type and QBD processes. Leeman [87] uses the same approach to derive the complementary distribution of stationary waiting times in more complex queuing system. We adopt their approach to derive the distribution of response time of low priority customers.

Let us tag a low priority customer entering the system. The time spent by this tagged customer depends on the number of customers of either class already present in the system ahead of it and also on the number of high priority arrivals before this tagged customers completes its service. All further low priority arrivals have no influence on its response time. Therefore, the time spent by this tagged customer in the system is the time until absorption in a modified Markov process $\{\widetilde{N}\}$, obtained by setting $\lambda_l = 0$. Consequently, the matrix A_1, representing transition to a higher level, becomes a zero matrix. Furthermore, we define an absorbing state 0^* as the state in which the tagged customer has finished its service and exits the system. The generator for this process is as follows:

$$\widetilde{Q} = \left(\begin{array}{c|ccccc} 0 & 0 & & & & \\ \hline b_0 & \widetilde{B_0} & 0 & & & \\ & A_{-1} & \widetilde{A_0} & 0 & & \\ & & A_{-1} & \widetilde{A_0} & 0 & \\ & & & \ddots & \ddots & \ddots \end{array} \right)$$

where, $\widetilde{B_0} = B_0 + A_1$; $\widetilde{A_0} = A_0 + A_1$, and the column vector b_0 contain elements that represent the transition from the state $(0, s)$ to the absorbing state 0^*. The first row and the column corresponds to the absorbing state $\widetilde{0}$.

The time spent by the tagged customer in the system is the time until absorption in the modified Markov process with rate matrix $\widetilde{Q}$. For a given arrival rates (λ_h and λ_l) and service rate μ, the distribution of the time spent by a low priority customer in the system is $S_j^l(\tau) = 1 - \overline{S_j^l}(\tau)$, where $\overline{S_j^l}(\tau)$ is the stationary probability that a low priority customer spends *more than* τ units of time in the system. Further, let $\overline{S_{jk}^l}(\tau)$ denote the conditional probability that the a tagged customer, who finds k low priority customers ahead of it, spends more than τ units of time in the system. The probability that the tagged customer finds m low priority customers ahead of it is given by $\pi = \pi_0 R^m$. Using the law of total probability, it follows that

$$\overline{S_j^l}(\tau) = \sum_{m=0}^{\infty} \pi_m \overline{S_{jm}^l}(\tau)$$

$\overline{S_{jk}^l}(\tau)$ can be computed by *uniformizing* the CTMC with a Poisson process with rate θ, where $\theta = \max_{0 \le m \le M}(-\widetilde{Q})_{mm} = \max_{0 \le i \le M}(-\widetilde{A_0})_{mm} = \max_{0 \le i \le M}(-A_1 - A_0)_{mm}$ so that the rate matrix $\widetilde{Q}$ is transformed into a discrete-time probability matrix:

$$\widehat{Q} = \frac{1}{\theta}\widetilde{Q} + I = \left(\begin{array}{c|cccc} 1 & 0 & & & \\ \hline \widehat{b_0} & \widehat{B_0} & & & \\ & \widehat{A_{-1}} & \widehat{A_0} & & \\ & & \widehat{A_{-1}} & \widehat{A_0} & \\ & & & \ddots & \ddots & \ddots \end{array}\right)$$

where $\widehat{b_0} = b_0/\theta$, $\widehat{B_0} =$, $\widehat{A_0} = \frac{\tilde{A}}{\theta} + I$, and $\widehat{A_{-1}} = A_{-1}/\theta$. In this uniformized process, the points of a Poisson process are generated with rate θ and transitions occur at these epochs only. The probability that t Poisson points are generated in time τ equals $e^{-\theta\tau}\frac{(\theta\tau)^t}{t!}$. Suppose that the tagged customer finds m customers ahead of it. Then, for the time in system to exceed τ, at most m of the t Poisson points may correspond to transitions to lower leel (). Therefore

$$\overline{S^l_{jm}}(\tau) = \sum_{t=0}^{\infty} e^{-\theta\tau}\frac{(\theta\tau)^t}{t!}\sum_{p=0}^{m} G_p^{(t)}\mathbf{e}$$

where $G_p^{(t)}$ is a matrix of conditional probabilities, given that the system has made t transitions in the discrete-time Makov process with rate matrix $\widehat{Q}$, and p of those transitions correspond to lower levels (i.e. service completions of low priority customers). The expression $\overline{S^l_j}(\tau)$ is given by:

$$\overline{S^l_j}(\tau) = \sum_{m=0}^{\infty}\pi_m\overline{S^l_{jm}}(\tau) = \sum_{m=0}^{\infty}\pi_0 R^m\sum_{t=0}^{\infty} e^{-\theta\tau}\frac{(\theta\tau)^t}{t!}\sum_{p=0}^{m} G_p^{(n)}\mathbf{e}$$

Therefore, the expression for service level reduces to

$$S^l_j(\tau) = 1 - \overline{S^l_j}(\tau) = 1 - \sum_{m=0}^{\infty} e^{-\theta\tau}\frac{(\theta\tau)^t}{t!}\pi_0(I-R)^{-1}H_t\mathbf{e} \tag{4.15}$$

where the matrix $H_t = \sum_{p=0}^{n} R^p G_p^{(n)}$ and can be computed recursively as:

$$H_t = H_{(t-1)}\widehat{A_0} + RH_{(t-1)}\widehat{A_{-1}}$$

starting with $H_{(0)} = I$.

4.5 Solution Procedure

The linear model $[MC_L]$ with infinite number of constraints is amenable to an iterative cutting plane method, where the service level and its subgradients are estimated using either simulation or matrix analytic method. It differs from the traditional description of the algorithm (presented in the previous chapters) only in that we use either simulation or matrix analytic method to evaluate the service level function and its subgradients due to the lack of existence of an algebraic expression for the function.

The iterative use of simulation and mathematical programming in the context of cutting plane algorithm to deal with the problem of lack of closed form solution of performance measures of interest has drawn the attention of researchers very recently (see [13, 14, 38, 73, 98] and references therein). The idea of combining simulation and mathematical programming in an optimization framework seems very promising as it harnesses the advantages of two powerful solution techniques. Henderson and Mason [73] is the first to outline a general methodology that uses simulation with integer programming iteratively for solving rostering problems in call centers, where one wishes to minimize the costs of staffing, subject to the service level constraint. Morito et al. [98] use simulation in a cutting plane algorithm to solve a large-scale logistic system design problem at the Japanese Postal Service. Atlason et al. [13, 14] present an iterative simulation-based cutting plane algorithm to optimize the scheduling of agents in the context of call center staffing problem (with single-call-type and single-skill type) with the objective of minimizing total staffing costs subject to service-level requirements over multiple time periods. Cezik and L'Ecuyer [38] extend the methodology to optimize the scheduling of agents in the multiskill call center staffing problem.

We present an iterative simulation-based cutting plane algorithm to optimize the demand allocation and location and capacity acquisition decisions in the design of an MTO supply chain network comprising of spatially distributed service facilities (DCs) that would serve the demand of multiple classes of customers with prespecified service-level requirements. The idea is to optimize a relaxed version of the problem by generating cuts from the violated service-level constraints and adding corresponding linear constraints until the optimal solution of the relaxed problem is feasible for the original problem. For that, we relax the service-level constraints (5.13), and solve the linear MIP model to obtain an initial solution (x^0, y^0). Using the demand allocation and the capacity level at the DCs, we compute the aggregate arrival rates (λ_j^n) and service rates (μ_j) at all the DCs selected open. We run simulation with the arrival rates and service rates obtained from the solution to get the estimates of service level function S_j^l and its three subgradients. If these estimates satisfy the service-level constraints (5.13), then we stop with the optimal solution to model $[MC]$, else we add a set of linear constraints of the form (5.13) to the relaxed problem so that it will eliminate the current solution without eliminating any feasible solution. This procedure is repeated until all the service level constraints are satisfied. The convergence of this solution procedure can be proved along the lines of Atlason et al. [13, 14].

4.6 Computational Results and Insights

In this section, we report our computational experiences with the proposed solution procedure and present some insights. The proposed solution procedures were coded in C and the MIP problems were solved using ILOG CPLEX 10.1 (using the Callable Library). The simulation model was built in SimEvents. The matrix analytic method was coded in MATLAB. The tests were performed on a Sun Blade 2500 workstation with 1.6-GHz UltraSPARC IIIi processors. The test problems are generated as on procedure outlined in Chapter 2. In all the test problems, we consider two customer classes - the mean demand arrival rate of the high priority customer (λ_j^h) is obtained by dividing the population of the cities by 10^3. The mean demand arrival rate of the low priority customer (λ_j^l) is set to: $\lambda_j^l = U[0.5, 1.5] \times \lambda_j^h$. The service level requirements are set to $\alpha^h = 99\%, 98\%$, $\alpha^l = 90\%$ to 95%. The threshold on the waiting time τ^n is set to expected waiting time in an M/M/1 queue: $\min_j\{\lambda_j/\mu_j(\mu_j - \lambda_j) + 1/\mu_j\}$, where $\lambda_j = \lambda_j^h + \lambda_j^l$ and μ_j are the arrival and service rates of the DCs selected open at the first iteration of the solution procedure (i.e. based on initial solution - x^0, y^0).

In the implementation of the matrix analytic procedure, the number of labels of high and low priority customer classes are set to M = 100. Figure 4.2 shows the effect of changing the number of levels (M) on the distribution of response time for low priority class customers. The figure depicts that as the value of M increases, the error introduced due to the truncation decreases, hence we set M = 100. Figure 4.3 shows the comparison of the estimates of the response time distributions of the low priority class customers obtained from simulation and the matrix geometric method for different utilization levels ($\rho = 0.3, 0.6, 0.9$). The vertical lines show the 99% confidence interval estimates from simulation, whereas the dotted line shows the estimates from matrix geometric method. It is clear from this figure that both the estimates are very close.

The step size for estimating the subgradient using the finite difference method are set to: $d\lambda_h = d\lambda_l = d\mu = 0.05$. The number of replications required for the simulation is set to ensure that the estimates of the service level function and the subgradients are obtained at the 99% confidence intervals [85]. For example, if the desired level of confidence are set at 99%, then the confidence interval is given by $\left(\overline{\frac{\partial S_j^l}{\partial \lambda_j}} - h, \overline{\frac{\partial S_j^l}{\partial \lambda_j}} + h\right)$, where

$$h = t_{n-1,1-\alpha/2}\sqrt{\frac{\sum_{d=1}^{n}\left(\frac{\partial S_j^l}{\partial \lambda_j}\right)_d^2 - n\left(\overline{\frac{\partial S_j^l}{\partial \lambda_j}}\right)^2}{n(n-1)}}$$

where t is constant obtained from the statistical tables depending on α and n; $\alpha = 0.01$, if 99% is the desired level of confidence in the estimate; n is the number of replications, $\overline{\frac{\partial S_j^l}{\partial \lambda_j}}$ is the mean subgradient estimate over n replications; and $\left(\frac{\partial S_j^l}{\partial \lambda_j}\right)_d$ is the subgradient estimate at replication d.

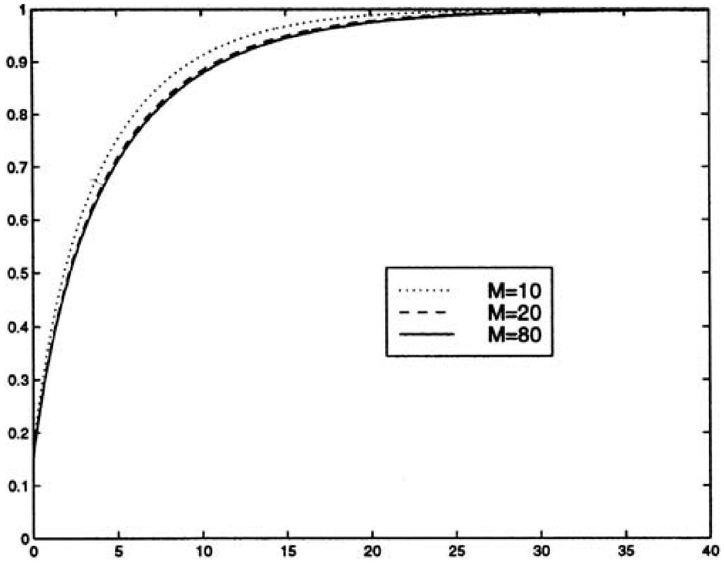


Figure 4.2: Effect of changing the number of levels (M) on the distribution of response time for low priority class customers (M/M/1 case)

In Table 4.1, we compare the performance of the simulation-based cutting plane method (S-CPM) and the matrix analytic based cutting plane method (MGM-

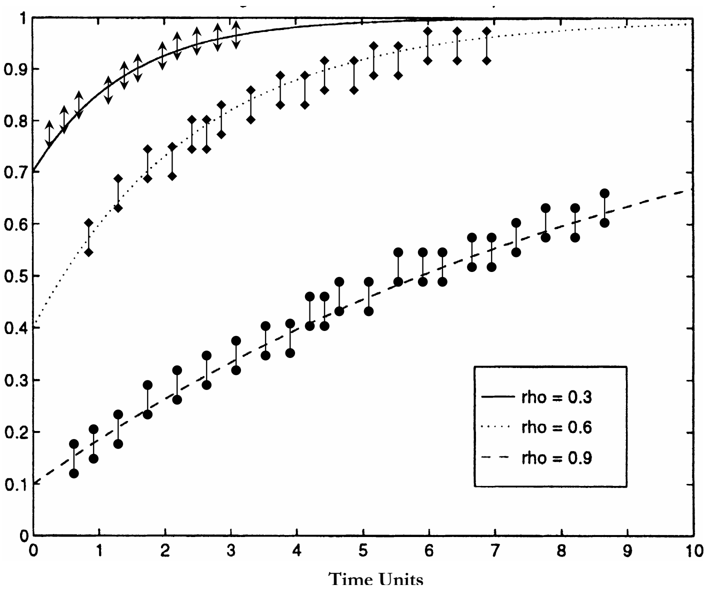


Figure 4.3: Comparison of the estimates of the response time distributions of the low priority class customers obtained from simulation and the matrix geometric method for different utilization (M/M/1 Case)

CPM) and report the results for eight test problems by varying the service level requirements, α^h and α^l for M/M/1 case. It is worthwhile nothing that simulation based cutting plane method can be used to deal with M/G/1 case provided the distribution of service processes is known and the concavity assumption for the performance measures of interest holds. However, the use of matrix analytic method is restricted to problems which can be modelled as quasi-birth-and-death processes (e.g. M/M/1 case).

In Table 4.1, the columns marked FC and VC represent the fixed costs and the variable production and transportation expressed as a percentage of total costs, DC represents the number of DCs opened. The table also displays the number of constraints generated (CUT), the number of iterations of the method (ITR), and the total CPU time in seconds required to obtain the optimal solution. The results show that both the solution procedures succeeded in finding the optimal solution to these test instances. However, the simulation-based cutting plane method outperforms the other method in terms of computational times - on average, S-CPM requires 285 sec, whereas the MAM-CPM requires 466 sec. This is due to the matrix computations required by the matrix analytic method for determining the service level. Furthermore, as the service level requirements increase, the methods require more iterations and computation time as large number of cuts are required. The results also show that increasing service level requirements reduces congestion by either increasing the capacity of the DCs, opening new DCs, or reallocating customer demand to among various DCs.

4.7 Concluding Remark

In this chapter, we presented a model for designing MTO supply chains for segmented markets with service-level differentiated customers. The model seeks to simultaneously determine the location and the capacity of the DCs, and allocate stochastic customer demand to DCs by minimizing the fixed location cost and the variable production and transportation cost subject to service level constraints for multiple demand classes. We presented a simulation-based cutting plane method, where we use simulation to estimate the service-level function and its subgradients. We compared the results with the matrix analytic method based cutting plane algorithm for M/M/1 case and found that simulation outperforms in terms of computational times. In future, we would like to explore the use of simulation-based cutting plane method in a Lagrangean framework for solving larger instances of the problem. In summary, the solution method looks promising and can be used to analyze more complex problems in supply chain optimization, for which no closed form expression for performance measures of interest exists.

Table 4.1: Comparison of the simulation-based cutting plane method and the matrix analytic method based cutting plane method: MTO supply chain design (M/M/1 Case - preemptive priority)

						Simulation-based Cutting Plane Method (S-CPM)						Matrix-Analytic-based Cutting Plane Method (MA-CPM)					
No.	I	J	K	α_h	α_l	FC	VC	DC	CUT	ITR	CPU	FC	VC	DC	CUT	ITR	CPU
						(%)	(%)				(s)	(%)	(%)				(s)
1	50	10	3	0.98	0.90	45	55	6	13	12	50	45	55	6	11	13	52
				0.98	0.91	45	55	6	13	15	63	45	55	6	11	17	59
				0.98	0.92	48	52	6	17	12	65	48	52	6	12	15	70
				0.99	0.93	48	52	6	18	12	69	48	52	6	12	16	91
				0.99	0.94	48	52	6	18	10	70	48	52	6	15	14	108
				0.99	0.95	48	52	6	23	18	78	48	52	6	15	11	123
2	50	20	3	0.98	0.90	60	40	11	21	18	116	60	40	11	20	23	272
				0.98	0.91	62	38	11	21	17	117	62	38	11	24	22	299
				0.98	0.92	62	38	11	21	14	120	62	38	11	24	25	307
				0.99	0.93	67	33	12	25	15	110	67	33	12	28	22	310
				0.99	0.94	67	33	12	25	15	137	67	33	12	32	20	346
				0.99	0.95	67	33	12	32	15	146	67	33	12	35	23	364
3	100	5	3	0.98	0.90	30	70	4	18	23	180	30	70	4	37	12	395
				0.98	0.91	32	68	4	22	23	182	32	68	4	37	17	417
				0.98	0.92	35	65	4	22	23	182	35	65	4	37	24	425
				0.99	0.93	40	60	5	22	27	189	40	60	5	38	21	433
				0.99	0.94	42	58	5	28	27	204	42	58	5	38	16	441
				0.99	0.95	43	57	5	28	26	213	43	57	5	38	17	446
4	100	10	3	0.98	0.90	48	52	8	27	12	374	48	52	8	25	13	457
				0.98	0.91	50	50	9	32	12	391	50	50	9	25	16	482
				0.98	0.92	53	47	9	28	12	386	53	47	9	25	15	487
				0.99	0.93	53	47	9	30	15	231	53	47	9	27	13	524
				0.99	0.94	54	46	9	30	21	408	54	46	9	27	18	525
				0.99	0.95	55	45	8	37	21	263	55	45	9	27	19	544
5	100	20	3	0.98	0.90	48	52	8	23	21	437	48	52	8	15	17	559
				0.98	0.91	50	50	8	23	21	550	50	50	8	15	15	561
				0.98	0.92	52	48	9	21	21	336	52	48	9	15	18	563
				0.99	0.93	52	48	9	29	24	404	52	48	9	21	10	564
				0.99	0.94	55	45	10	34	24	518	55	45	10	21	15	569
				0.99	0.95	56	44	10	25	27	375	56	44	10	21	15	575
6	150	5	5	0.98	0.90	30	70	4	15	20	212	30	70	4	23	10	422
				0.98	0.91	32	68	4	30	20	278	32	68	4	23	17	316
				0.98	0.92	35	65	4	23	21	297	35	65	4	23	18	417
				0.99	0.93	38	62	4	28	21	239	38	62	4	27	11	517
				0.99	0.94	40	60	5	25	23	322	40	60	5	27	17	468
				0.99	0.95	41	58	5	21	23	354	41	58	5	27	23	405
7	150	10	5	0.98	0.90	30	70	6	21	15	271	30	70	6	18	13	531
				0.98	0.91	32	68	6	17	15	211	32	68	6	22	18	533
				0.98	0.92	33	67	6	12	16	202	33	67	6	27	16	534
				0.99	0.93	35	62	7	16	17	234	35	62	7	32	17	574
				0.99	0.94	36	64	7	25	18	243	36	64	7	32	21	575
				0.99	0.95	38	56	7	13	21	278	38	56	7	32	21	598
8	150	20	5	0.98	0.90	49	51	10	27	21	533	49	51	10	36	16	756
				0.98	0.91	51	49	10	43	21	531	51	49	10	45	17	852
				0.98	0.92	52	48	10	45	21	544	52	48	10	45	18	993
				0.99	0.93	52	48	10	46	22	672	52	48	10	53	21	749
				0.99	0.94	52	48	10	47	22	582	52	48	10	53	21	872
				0.99	0.95	54	46	10	49	26	691	54	46	10	58	24	880
					Min	30	33	4	12	10	50	30	33	4	11	10	52
					Max	67	70	12	49	27	691	67	70	12	58	25	993
					Mean	47	53	8	26	19	285	47	53	8	28	17	466

I: No. of customers; J: No. of potential DCs; K: No of capacity levels at each DC;
α_h: specified service-level for high priority customers;
α_l: specified service-level for low priority customers;
FC: Fixed cost; VC: Variable cost (expressed as percentages of the total cost);
DC: No. of DCs selected open;
CUT: Number of cuts generated;
ITR: Number of iterations;
CPU(s): CPU time in sec.

Chapter 5

Conclusions and Future Research

We summarize the major contributions of this thesis. We also discuss several variations and extensions to current models and some additional research directions.

5.1 Summary

The major contribution of this thesis lies the modelling and development of solution approaches for the design of *responsive supply chain* networks - MTO and ATO supply chains. We present three models: (i) model for MTO supply chain design that attempt to find solutions that balances the overall response time cost and the fixed cost of DC location and capacity acquisition and the variable transportation cost, (ii) model for ATO supply chain design that attempt to find solutions that balances the overall response time cost and the fixed cost of plant and DC location and capacity acquisition, and the inbound and outbound transportation cost, and (iii) model for MTO supply chain design that attempt to find solutions that balances the response time based service level requirements for multiple customer classes against the fixed cost of DC location and capacity acquisition and the variable transportation cost. It is evident that inclusion of stochastic demand, response time considerations and service level requirements in presence of congestion and multiple customer classes made the model highly nonlinear and difficult to solve. To this end, we presented linearization procedures, exact and heuristic solution approaches for these models that performed well, providing optimal solutions, tighter bounds and short computation times. We show using numerical examples that large improvements in response time are often possible with little additional cost.

5.2 Future Research Directions

The summary presented above shows the current understanding of an integrated large-scale supply chain model. It is certainly true that there are more issues to be explored as the expansion of the knowledge of these systems continues and the realization of the benefits of integrated supply chain grows. There are several potential directions in which the research reported here can be extended. Some of them are as follows:

- In the design of MTO supply chains, we have so far modelled DCs as single server queues. This is appropriate only when DCs experience heavy traffic (i.e. $\rho_j \approx 1$). Under heavy traffic conditions, a DC with multiple (identical) servers will operate almost identical to a single server queue with service rate equal to the sum of all the parallel servers [33]. However, it is difficult to predict in advance whether all the DCs will experience heavy traffic, especially when the objective is to reduce the overall response time. Therefore, it would be interesting to look at cases where DCs are modelled as network of spatially distributed multiple server queues with Poisson demand arrivals and general service time distributions (M/G/s). Due to the lack of a closed form expression for waiting time distributions in a M/G/s queue, one might have to rely on simulation.

- It would be interesting to extend the model in Chapter 4 to consider the design of two-echelon ATO supply chains for segmented markets with service-level differentiated customers.

- One can investigate the combined effect of risk pooling and congestion on the design of response time sensitive supply chains. While the risk-pooling effect would consolidate the workloads on a few DCs, the congestion components would distribute the demand among various DCs to reduce the overall response time.

- Simultaneous location, capacity, and pricing in presence of congestion: In addition to the waiting time and access costs, customers are sensitive to price of the product. Therefore, it would be interesting to look at the interaction between the location, capacity, and pricing decisions of a firm using analytical models and solution approaches.

- The combined use of large-scale optimization and simulation can prove to be an efficient solution approach in the modelling and analysis of integrated supply chains. These approaches would make use of discrete-event simulation for the estimation of performance measures that cannot be evaluated using analytical models or mathematical programming models.

- We would also look into the possibility of using simulation in solving subproblems, while using optimization to deal with master problem in a Lagrangean relaxation framework. This can be useful for large-scale optimization problems arising in the design and planning of service parts logistics network design and (e.g. IBM Spare parts Division) and other express package delivery distribution system design (e.g. FedEx, UPS).

Bibliography

[1] G. Abate, G. L. Choudhury, and W. Whitt. Waiting time tail probabilities in queues with long-tail service time distributions. *Queueing Systems*, 16:311–338, 1994. 69

[2] R. Aboolian, O. Berman, and Z. Drezner. Location and allocation of service units on a congested network. *IIE Transactions*, 40:422–433, 2008. 17

[3] R. Aboolian, O. Berman, and Z. Drezner. The multiple server center location problem. *Annals of Operations Research*, 2008. 17

[4] R. Aboolian, O. Berman, and D. Krass. Competitive facility location model with concave demand. *European Journal of Operational Research*, 181:598–619, 2007. 33

[5] S. R. Agnihotri, S. Narasimhan, and H. Pirkul. Assignment problem with queuing time cost. *Naval Research Logistics*, 37:231–244, 1990. 16, 17

[6] S. L. Albin. Approximating a point process by a renewal process, ii: Superposition of arrival processes to queues. *Operations Research*, 32:1133–1162, 1984. 50

[7] G. Allon and A. Federgruen. Competition in service industries with segmented markets. *Working paper*, 2007. Kellogg School of Management, Northwestern University, IL. 6

[8] A. Amiri. Solution procedures for the service system design problem. *Computers and Operations Research*, 24:49–60, 1997. 17, 31

[9] A. Amiri. The design of service systems with queuing time cost, workload capacities, and backup service. *European Journal of Operational Research*, 104:201–217, 1998. 17

[10] A. Amiri. The multi-hour service system design problem. *European Journal of Operational Research*, 128:625–638, 2001. 17

[11] S. Andradottir. Simulation optimization. In J. Banks, editor, *Handbook of Simulation*, pages 307–333, New York, New York, 1998. John Wiley and Sons. 78, 79

[12] B. C. Arntzen, C. G. Brown, T. P. Harrison, and L. Trafton. Global supply chain management at Digital Equipment Corporation. *Interfaces*, 25(1):69–93, 1995. 14

[13] J. Atlason, M. A. Epelman, and S. G. Henderson. Call center staffing with simulation and cutting plane methods. *Annals of Operations Research*, 127:333–358, 2004. 78, 86, 87

[14] J. Atlason, M. A. Epelman, and S. G. Henderson. Optimizing call center staffing with simulation and analytic center cutting-plane methods. *Management Science*, 54(2):295–309, 2008. 78, 86, 87

[15] R. H. Ballou. DISPLAN: A miltiproduct plant/warehouse location model with non-linear inventory costs. *Journal of Operations Management*, 5(1):75–91, 1984. 15

[16] F. Barahona and D. Jensen. Plant location with minimun inventory. *Mathematical Programming*, 83:101–111, 1998. 15

[17] O. Baron, O. Berman, and D. Krass. Facility location with stochastic demand and constraints on waiting time. *Manufacturing and Service Operations Management*, 10(3):484–505, 2008. 17, 19

[18] R. Batta. A queuing-location model with expected service time dependent queuing disciplines. *European Journal of Operational Research*, 39:192–205, 1989. 71

[19] R. Batta, R. C. Larson, and R. Odoni. A single-server priority queueing-location model. *Networks*, 8:87–103, 1988. 71

[20] S. Benjaafar and M. ElHafsi. Production and inventory control of a single product assemble-to-order system with multiple customer classes. *Management Science*, 52(12):1896–1912, 2006. 7

[21] S. Benjaafar, M. ElHafsi, and F. Vericourt. Demand allocation in multi-product, multi-facility, make-to-stock systems. *Management Science*, 50(10):1431–1448, 2004. 50

[22] S. Benjaafar, Y. Li, D. Xu, and S. Elhedhli. Demand allocation in systems with multiple inventory locations and multiple demand sources. *Manufacturing and Service Operations Management*, 10(1):43–60, 2008. 50, 57

[23] O. Berman and Z. Drezner. Location of congested capacitated facilities with distance sensitive demand. *IIE Transactions*, 38:213–221, 2006. 17

[24] O. Berman and Z. Drezner. The multiple server location problem. *Journal of the Operational Research Society*, 58:91–99, 2007. 17

[25] O. Berman and D. Krass. Facility location problems with stochastic demands and congestion. In Z. Drezner and H.W. Hamacher, editors, *Facility Location: Applications and Theory*, pages 329–371. Springer, 2004. 17

[26] O. Berman, D. Krass, and J. Wang. Locating service facilities to reduce lost demand. *IIE Transactions*, 38:933–946, 2006. 17

[27] O. Berman, R. C. Larson, and S. S. Chiu. Optimal server location on a network operating as an M/G/1 queue. *Operations Research*, 12:746–771, 1985. 17

[28] J. D. Blackburn. *Time-based Competition: The Next Battle Ground in American Manufacturing*. Richard D. Irwin, Boston, 1991. 1, 2, 12

[29] B. Boffey, R. Galvao, and L. Espejo. A review of congestion models in the location of facilities with immobile servers. *European Journal of Operational Research*, 178(3):643–662, 2007. 17, 19

[30] J. H. Bookbinder and K. E. Reece. Vehicle routing considerations in distribution system design. *European Journal of Operations Research*, 37:204–213, 1988. 15

[31] G. Brown, J. Keegan, B. Vigus, and K. Wood. The Kellogg company optimizes production, inventory, and distribution. *Interfaces*, 31(6):1–15, 2001. 14

[32] G. G. Brown, G. W. Graves, and M. D. Honczarenko. Design and operation of a multicommodity production/distribution system using primal goal decomposition. *Management Science*, (33):1469–1480, 1987. 14

[33] J. A. Buzacott and J. G. Shanthikumar. *Stochastic Models of Manufacturing System.* Prentice-Hall, Engelwood Cliffs, New Jersey, 1993. 50, 57, 93

[34] J. D. Camm, T. E. Chorman, F. A. Dill, J. R. Eans, D. J. Sweeney, and G. W. Wegryn. Blending OR/MS, judgement, and GIS: Restructuring P & G's supply chain. *Interfaces*, 27(1):128–142, 1997. 14

[35] M. F. Candas and E. Kutanoglu. Benefits of considering inventory in service parts logistics network design problems with time-based service level constraints. *IIE Transactions*, 39:159–176, 2007. 15

[36] C. Canel and B. M. Khumawala. Multi-period international facilities location: An algorithm and application. *International Journal of Production Research*, 35(7):1891–1910, 1997. 15

[37] I. Castillo, A.Ingolfsson, and S. Thaddues. Socially optimal location of facilities with fixed servers, stochastic demand, and congestion. *Working paper.* School of Business, University of Alberta. 17

[38] M. T. Cezik and P. L'Ecuyer. Staffing multiskill call centers via linear programming and simulation. *Management Science*, 54(2):310–323, 2008. 78, 86

[39] H. Chen and S. G. Henderson. Two issues in setting call center staffing levels. *Annals of Operations Research*, 108(1):175–192. 75

[40] S. Choi, A. Seidmann, and M. W. Suh. Decision models for designing and planning private communication networks. *Decision Support Systems*, pages 389–403, 1995. 16

[41] S. Chopra and P. Meindl. *Supply Chain Management: Strategy, Planning, and Operation.* Pearson Education, Upper Saddle River, New Jersey, 2001. 8, 13

[42] M. A. Cohen and H. L. Lee. Resource deployment analysis of global manufacturing and distribution networks. *Journal of Manufacturing and Operations Management*, 2:81–104, 1989. 15

[43] J.-F. Cordeau, F. Pasin, and M. M. Solomon. An integrated model for logistics network design. *Annals of Operations Research*, 144:59–82, 2006. 15

[44] K. L. Croxton and W. Zinn. Inventory considerations in network design. *Journal of Business Logistics*, 26(1):149–168, 2005. 15

[45] A. Dasci and V. Verter. The plant location and technology acquisition problem. *IIE Transactions*, 33:963–973, 2001. 15

[46] M. Daskin, L. Snyder, and R. Berger. Facility location models in supply chain design. In A. Langevin and D. Riopel, editors, *Logistic Systems: Design and Optimization*, pages 39–65. Kluwer Academic Publishers, 2005. 15

[47] M. S. Daskin. *Network and Discrete Location: Models, Algorithms, and Applications.* Wiley-Interscience, New York, 1995. 32

[48] M. S. Daskin. SITATION- Facility location software. Department of Industrial Engineering and Management Sciences, Northwestern Univeristy, Evanston, IL. Available at http://users.iems.nwu.edu/∼msdaskin/. 2004. 63

[49] M. S. Daskin, C. R. Coullard, and Z.-J. Max Shen. An inventory location model: Formulation, solution algorithm and computational results. *Annals of Operations Research*, 110:83–106, 2002. 15

[50] M. Dell. *Direct from Dell: Strategies that Revolutionized an Industry.* Harper Collins, New York, 2000. 4

[51] V. Deshpande, M. A. Cohen, and K. Donohue. An empirical study of service differentiation for weapon system service parts. *Operations Research*, 51(4):518–530, 2003. 68

[52] V. Deshpande, M. A. Cohen, and K. Donohue. A threshold inventory rationing policy for service-differentiated demand classes. *Management Science*, 49(6):683–703, 2003. 70

[53] K. Dogan and M. Goetschalkx. A primal decomposition method for the integrated design of multi-period production-distribution systems. *IIE Transactions*, 31(11):1027–1036, 1999. 15

[54] M. A. Duran and I. E. Grossman. An outer approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical Programming*, 36(1):307–339, 1986. 23

[55] S. Duran, T. Liu., D. Simchi-Levi, and J. L. Swann. Optimal production and inventory policies for priority and price-differentiated customers. *IIE Transactions*, 39:845–861, 2007. 70

[56] S. Elhedhli. Exact solution of a class of nonlinear knapsack problems. *Operations Research Letters*, 33:615–624, 2005. 19, 33

[57] S. Elhedhli. Service system design with immobile servers, stochastic demand and congestion. *Manufacturing and Service Operations Management*, 8(1):92–97, 2006. 16, 17, 31

[58] S. Elhedhli and F. Gzara. Integrated design of supply chain networks with three echelons, multiple commodities, and technology selection. *IIE Transactions*, 40(1):31–44, 2008. 15

[59] S. S. Erengüç, N. C. Simpson, and A. J. Vakharia. Integrated production/distribution planning in supply chains: An invited review. *European Journal of Operational Research*, 115:219–236, 1999. 15, 16

[60] S. J. Erlebacher and R. D. Meller. The interaction of location and inventory in designing distribution systems. *IIE Transactions*, 32:155–166, 2000. 15

[61] E. Eskigun, R. Uzsoy, P. V. Preckel, G. Beaujon, S. Krishnan, and J. D. Tew. Outbound supply chain network design with mode selection, lead times, and capacitated vehicle distribution centers. *European Journal of Operational Research*, 165(1):182–206, 2005. 15, 22

[62] E. Eskigun, R. Uzsoy, P. V. Preckel, G. Beaujon, S. Krishnan, and J. D. Tew. Outbound supply chain network design with mode selection and lead times consideration. *Naval Research Logistics*, 54:282–300, 2007. 15

[63] M. L. Fisher. What is the right supply chain for your product? *Harvard Business Review*, pages 105–116, March-April 1997. 3

[64] B. Fleischmann, S. Ferber, and P. Henrich. Strategic planning of BMW's global production network. *Interfaces*, 36(3):194–208, 2006. 14

[65] K. C. Frank, R. Q. Zhang, and I. Duenyas. Optimal policies for inventory systems with priority demand classes. *Management Science*, 51(6):993–1002, 2003. 70

[66] K. Fridgeirsdottir and S. Chiu. A note on convexity of the expected delay cost in single-server queues. *Operations Research*, 53(3):568–570. 51

[67] A. Geoffrion and G. Graves. Multicommodity distribution system design by Benders decomposition. *Management Science*, (29):822–844, 1974. 14

[68] P. Glasserman and Y. Wang. Leadtime-inventory tradeoffs in assemble-to-order systems. *Operations Research*, 46(6):858–871, 1998. 13

[69] D. Gupta and S. Benjaafar. Make-to-order, make-to-stock, or delay product differentiation? a common framework for modelling and analysis. *IIE Transactions*, 36:529–546, 2004. 4, 8

[70] A. Y. Ha. Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science*, 43(8):1093–1103, 1997a. 70

[71] A. Y. Ha. Stock-rationing policy in a make-to-stock production system with two-priority classes and backordering. *Naval Research Logistics*, 44(8):457–472, 1997b. 70

[72] A. Y. Ha. Stock-rationing policy in an $M/E_k/1$ make-to-stock queue. *Management Science*, 46(1):77–87, 2000. 70

[73] S. G. Henderson and A. J. Mason. Rostering by integrating integer programming and simulation. In D. Medeiros, E. Watson, J. S. Carson, and M. S. Manivannan, editors, *Proceedings of the 1998 Winter Simulation Conference*, pages 677–684, Piscataway, NJ: IEEE, 1998. 86

[74] Y. Hinojosa, J. Kalcsics, S. Nickel, J. Puerto, and S. Velten. Dynamic supply chain design with inventory. *Computers and Operations Research*, 35(2):373–391, 2008. 15

[75] S. Huang, R. Batta, and R. Nagi. Distribution network design: selection and sizing of congested connections. *Naval Research Logistics*, 52:701–712, 2005. 15, 16, 55

[76] E. P. Robinson Jr., S. D. Muggenborg, and L. Gao. Designing an integrated distribution system at Dowbrands Inc. *Interfaces*, 23(3):107–117. 14

[77] N. Karabakal, A. Günal, and W. Ritchie. Supply-chain analysis at Volkswagen of America. *Interfaces*, 30(4):46–55, 2000. 14

[78] J. E. Kelley. The cutting plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8(4):703–712, 1960. 29, 61

[79] I. Kim and C. S. Tang. Lead time and response time in a pull production control system. *European Journal of Operational Research*, 101:474–485, 1997. 1, 2, 30, 31

[80] S. Kim and R. Uzsoy. Exact and heuristic procedures for capacity expansion problems with congestion. *IIE Transactions*, 40:1185–1197, 2008. 14

[81] A. Klose and A. Drexl. Facility location models for distribution system design. *European Journal of Operational Research*, 162:4–29, 2005. 15, 16, 58

[82] P. Kouvelis, M. J. Rosenblatt, and C. L. Munson. A mathematical programming model for global supply plant location problems: Analysis and insights. *IIE Transactions*, 36:127–144, 2004. 15

[83] G. Latouche and V. Ramaswami. *An Introduction to Matrix Analytic Methods in Stochastic Modeling*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1999. 77, 79

[84] C. Laval, M. Feyhl, and S. Kakouros. Hewlett-Packard combined OR and expert knowledge to design its supply chains. *Interfaces*, 35:238–247, 2005. 14

[85] A. M. Law and W. D. Kelton. *Simulation Modelling and Analysis*. Mcgraw Hill, Boston, 2000. 78, 88

[86] H. L. Lee and C. Billington. Supply chain management: Pitfalls and opportunities. *Sloan Management Review*, 33:65–73, 1992. 8

[87] H. T. Leeman. Waiting time distribution in a two-class two-server heterogeneous priority queues. *Performance Evaluation*, 43:133–150, 2001. 83

[88] H. Luss. Operations research and capacity expansion problems: A survey. *Operations Research*, 30(5):907–947, 1982. 13

[89] J. Margretta. The power of virtual integration: An interview with Dell Computer's Michael Dell. *Harvard Business Review*, 76(2), March-April 1998. 2, 4

[90] V. Marianov and D. Serra. Location-allocation of multiple-server service centers with constrained queues or waiting times. *Annals of Operations Research*, 111:35–50, 2002. 17

[91] A. Martel. The design of production-distribution networks: A mathematical programming approach. In J. Geunes and P. M. Pardalos, editors, *Supply Chain Optimization*. Springer. 15

[92] C. H. Martin, D. C. Dent, and J. C. Eckhart. Integrated production, distribution and inventory planning at Libbey-Owens-Ford. *Interfaces*, 23(3):68–78, 1993. 14

[93] M. J. Meixell and V. B. Gargeya. Global supply chain design: A literature review. *Transportation Research, Part E*, 41(6):531–550, 2005. 15

[94] P. Mellacheruvu, M. C. Fu, and J. W. Herrmann. Comparing gradient estimation methods applied to stochastic manufacturing systems. *Technical Research Report T.R.2000-1, Institute for Systems Research, University of Maryland*, 101:474–485, 2000. 79

[95] P. A. Miranda and R. A. Garrido. Incorporating inventory control decisions into a strategic distribution network design model with stochastic demand. *Transportation Research, Part-E*, 40(3):183–207, 2004. 15

[96] P. A. Miranda and R. A. Garrido. Valid inequalities for Lagrangean relaxation in an inventory location problem with stochastic capacity. *Transportation Research, Part-E*, 44:47–65, 2008. 15

[97] Y. Mo and T. P. Harrison. A conceptual framework for robust supply chain design under demand uncertainty. In J. Geunes and P. M. Pardolas, editors, *Supply Chain Optimization*. Kluwer Academic Publishers. 15

[98] S. Morito, J. Koida, T. Iwama, M. Sato, and Y. Tamura. Simulation-based constraint generation with applications to logistic system design. In P. A. Farrington, H. B. Nembard, D. T. Sturrock, and G. W. Ewans, editors, *Proceedings of the 1999 Winter Simulation Conference*, pages 531–536, 1999. 86

[99] F. M. Neuts. *Matrix Geometric Solutions in Stochastic Methods: An Algorithmic Approach*. Dover Publications, Mineola, NY, USA, 1981. 77, 79

[100] L. K. Nozick and M. A. Turnquist. Integrating inventory impacts into a fixed-charge model for locating DCs. *Transportation Research Part E*, 34:173–186, 1998. 15

[101] L. K. Nozick and M. A. Turnquist. Inventory, transportation, service quality, and the location of distribution centers. *European Journal of Operational Research*, 129:362–371, 2001. 15

[102] L. K. Nozick and M. A. Turnquist. A two-echelon allocation and distribution center location analysis. *Transportation Research Part E*, 37:425–441, 2001. 15

[103] L. Ozsen, C. R. Coullard, and M. S. Daskin. Capacitated warehouse location model with risk pooling. *Naval Research Logistics*, 55(4):295–312, 2008. 15

[104] M. Paquet, A. Martel, and G. Desaulniers. Including technology selection decisions in manufacturing network design models. *International Journal of Computer Integrated Manufacturing*, 17(2):117–125, 2004. 15

[105] W. P. Peterson. A heay traffic limit theorem for networks of queues with multiple customer types. *Mathematics of Operations Research*, 16:90–118, 1991. 51

[106] J. Pooley. Integrated production and distribution facility planning at Ault Foods. *Interfaces*, 24(4):113–121, 1994. 14

[107] L. Qi and Z.-J. M. Shen. A supply chain design model with unreliable supply. *Naval Research Logistics*, 54:829–844, 2007. 15

[108] S. Rajagopalan and H.-L. Yu. Capacity planning with congestion effects. *European Journal of Operational Research*, 134:365–377, 2001. 14

[109] V. Ramaswami and D. M. Lucantoni. Stationary waiting time distribution in queues with phase type service and quasi-birth-and-death processes. *Communication in Statistics- Stochastic Models*, 1(2):125–136, 1985. 83

[110] U. Rao, A. Scheller-Wolf, and S. Tayur. Development of a rapid-response supply chain at Caterpillar. *Operations Research*, 48(2):189 –204, 2000. 14, 22

[111] S. Ray, Y. Gerchak, and E. M. Jewkes. The effectiveness of investment in lead time reduction for a make-to-stock product. *IIE Transactions*, 36:333–344, 2004. 13

[112] S. Ray and E. M. Jewkes. Customer lead time management when both demand and price are lead time sensitive. *European Journal of Operational Research*, 153:769–781, 2004. 13

[113] A. G. Robinson and J. H. Bookbinder. NAFTA supply chains: Facilities location and logistics. *International Transactions in Operational Research*, 14(2):179–199, 2007. 15

[114] H. E. Romeijn, J. Shu, and C.-P. Teo. Designing two-echelon supply networks. *European Journal of Operational Research*, 178:449–462, 2007. 15

[115] G. Ruderman. The state of automotive make-to-order; poor demand picture and legacy delay progress towards custom configurations. *Manufacturing Business Technology, Reed Business Information*, 22(8), August 2004 2004. 5

[116] A. M. Sarmiento and R. Nagi. A review of integrated analysis of production-distribution systems. *IIE Transactions*, 31:1061–1074, 1999. 15, 16

[117] Y. Sheffi. Creating demand responsive supply chains. *Harvard Business Review*, pages 3–5, April 2005. 3, 5

[118] Z.-J. M. Shen. A multi-commodity supply chain design problem. *IIE Transactions*, 37(8):753–762, 2005. 15

[119] Z.-J. M. Shen. A profit-maximizing supply chain network design model with demand choice flexibility. *Operations Research Letters*, 34:673–682, 2006. 15

[120] Z.-J. M. Shen. Integrated stochastic supply chain design. *Computing in Science and Engineering*, 9(2):50–59, 2007. 15

[121] Z.-J. M. Shen. Integrated supply chain design models: A survey and future research directions. *Journal of Industrial and Management Optimization*, 3(1):1–27, 2007. 15

[122] Z.-J. M. Shen, C. Coullard, and M. S. Daskin. A joint location inventory model. *Transportation Science*, 37(1):40–55, 2003. 15

[123] Z.-J. M. Shen and M. S. Daskin. Trade-offs between customer service and costs in integrated supply chain design. *Manufacturing and Service Operations Management*, 7(3):188–207, 2005. 15

[124] Z.-J. M. Shen and L. Qi. Incorporating inventory and routing costs in strategic location models. *European Journal of Operational Research*, 179:372–389, 2007. 15

[125] J. Shu and J. Sun. Designing the distribution network for an integrated supply chain. *Journal of Industrial and Management Optimization*, 2(3):339–349, 2006. 15

[126] F. Silva and D. Serra. Locating emergency services with different priorities: The priority queuing covering location problem. *Journal of Operational Research Society*, 59:1229–1238, 2008. 69, 71, 73

[127] D. Simchi-Levi, P. Kaminsky, and E. Simchi-Levi. *Designing and Managing the Supply Chain*. McGraw-Hill Irwin, New York, NY, 2003. 1, 5

[128] R. L. Simison. Toyota develops a way to make a car within five days of a customer order. *The Wall Street Journal*, page A.4, Aug 6, 1999. 5

[129] L. V. Snyder and M. Daskin. Models for reliable supply chain network design. In A. T. Murray and T. Grubesic, editors, *Critical Infrastructure: Reliability and Vulnerability*, pages 257–289. Berlin: Springer Verlag, 2007. 15

[130] L. V. Snyder and M. S. Daskin. Reliability models for facility location: The expected failure cost case. *Transportation Science*, 39:400–416, 2005a. 15

[131] L. V. Snyder and M. S. Daskin. Stochastic p-robust location problems. *IIE Transactions*, 38(11):971–985, 2006. 15

[132] K. C. So and J.-S. Song. Price, delivery time guarantees and capacity selection. *European Journal of Operational Research*, 111:28–49, 1998. 6, 13

[133] J.-S. Song and P. Zipkin. Supply chain operations: Assemble-to-order systems. In A. G. De Kok and S. Graves, editors, *Handbooks in Operations Research and Management Science, Vol. 11: Supply Chain Management: Design, Coordination, and Operation*. 4, 8, 53, 54

[134] K. Sourirajan, L. Ozsen, and R. Uzsoy. A single-product network design model with lead time and safety stock considerations. *IIE Transactions*, 39:411–424, 2007. 15, 16

[135] G. Jr. Stalk. Time - the next source of competitive advantage. *Harvard Business Review*, pages 41–51, July-August 1988. 1, 2

[136] G. Jr. Stalk and T. M. Hout. *Competing Against Time*. The Free Press, 1990. 1, 2, 6, 12

[137] J. S. Swaminathan. Enabling customization using standardized operations. *California Management Review*, 43(3):125–135, 2001. 4, 5

[138] C.-P. Teo, J. Ou, and M. Goh. Impact of inventory costs with consolidation of distribution centers. *IIE Transactions*, 52(33):99–110, 2001. 15

[139] C.-P. Teo and J. Shu. Warehouse-retailer network design problem. *Operations Research*, 52(3):396–408, 2004. 15

[140] J. B. Treece. Ford: Alex Trotman's daring global strategy. *Business Week*. 5

[141] H. Y. Tu and H. Kumin. A convexity result for a class of GI/G/1 queuing systems. *Operations Research*, 31(5):948–950. 22

[142] V. Verter and A. Dasci. The plant location and flexible technology acquisition problem. *European Journal of Operational Research*, 136:366–382, 2002. 15

[143] C. J. Vidal and M. Goetschalckx. Strategic production-distribution models: A critical review with emphasis on global supply chain models. *European Journal of Operational Research*, 98:1–18, 1997. 15, 16

[144] C. J. Vidal and M. Goetschalckx. Modeling the effect of uncertainities on global logistics systems. *Journal of Business Logistics*, 21(1):95–120, 2000. 15

[145] N. K. Vidyarthi, E. Çelebi, S. Elhedhli, and E. M. Jewkes. Integrated production-inventory-distribution system design with risk pooling: Model formulation and heuristic solution. *Transportation Science*, 41(3):392–408, 2007. 15

[146] N. K. Vidyarthi, S. Elhedhli, and E. M. Jewkes. Demand allocation and capacity decisions in make-to-order supply chain design for segmented markets with service-level differentiated customers. *Working paper*, (Department of Decision Sciences and MIS, John Molson School of Business, Concordia University, Montreal, Canada), 2009. 68

[147] N. K. Vidyarthi, S. Elhedhli, and E. M. Jewkes. Response time reduction in make-to-order and assemble-to-order supply chain design. *IIE Transactions*, 41(5):448–466, 2009. 9, 10

[148] Q. Wang, R. Batta, and C. M. Rump. Algorithms for a facility location problem with stochastic customer demand and immobile servers. *Annals of Operations Research*, 111:17–34, 2002. 17

[149] Q. Wang, R. Batta, and C. M. Rump. Facility location models for immobile servers with stochastic demand. *Naval Research Logistics*, 51:138–152, 2004. 17

[150] Y. Wang, M. A. Cohen, and Y.-S. Zheng. Differentiating customer service on the basis of delivery lead-times. *IIE Transactions*, 34:979–989, 2002. 70

[151] R. R. Weber. A note on waiting time in single server queues. *Operations Research*, 31(5):950–951. 22

[152] W. Whitt. Approximating a point process by a renewal process, ii: Two basic approaches. *Operations Research*, 30:125–147, 1982. 50, 51

[153] W. Wilhelm, D. Liang, B. Rao, D. Warrier, X. Zhu, and S. Bulusu. Design of international assembly systems and their supply chains under NAFTA. *Transportation Research - Part E*, 41:467–493, 2005. 15

[154] R. W. Wolff. *Stochastic modelling and the theory of queues*. Prentice-Hall, Engelwood Cliffs, New Jersey, 1989. 50, 75

[155] Y. Zhang, O. Berman, and V. Verter. Incorporating congestion in preventive healthcare facility network design. *European Journal of Operational Research*, 198:922–935, 2009. 17