Applications of Operations Management in Emerging Technologies

Mohammad Ebrahim Arbabian

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Kamran Moinzadeh, Chair

Shi Chen, Chair

Michael Wagner

Program Authorized to Offer Degree:
Foster School of Business

University of Washington

Abstract

Applications of Operations Management in Emerging Technologies

Mohammad Ebrahim Arbabian

Co-Chairs of the Supervisory Committee:
Kamran Moinzadeh
Information Systems and Operations Management

Shi Chen
Information Systems and Operations Management

The question of how to match supply and demand, in a supply chain, is crucial because it significantly affects the supply chain's cost management. Therefore, to adopt new technologies, supply chain owners should update their matching supply-demand policies. A common objective of my dissertation is to analyze and develop game-theoretic and stochastic models to propose new policies for supply chains whose supply or demand is influenced by emerging technologies. In particular, in the third chapter, after the introduction and literature review in the first two chapters, with a specific focus on cloud industry, we study markets in which one unit of supply increases the capacity of multiple attributes simultaneously. Working within this type of market can be problematic: since the ratio of the capacities of the attributes in one unit of supply differs from the attributes' demand ratio, balancing supply and demand is challenging and requires new policies. In the fourth chapter, we specifically focus on 3D printing, and investigate how it affects supply chains responsiveness to demand. Due to the unique features of 3D printing technology, 3D printers can be installed in a retail store in order for the retailer to produce a product if he faces demand shortage. Therefore, this technology, can potentially decrease the expected mismatch between supply and demand.

We, first, study a capacity expansion problem in cloud industry where the supply of attributes is bundled. The recent surge in demand for cloud services has posed a capacity expansion problem for service providers: while the growth of demand for capacity attributes (e.g., CPU and RAM) are time-varying and disproportionate, replenishments of these attributes are often in pre-configured packages (e.g., server clusters). This problem was originally introduced to us by Microsoft; however, our communications with other cloud companies such as Amazon and eBay indicate that they face similar problems. In the third chapter, we consider demand growths of two attributes (CPU and RAM) and focus on a class of policies consisting of capacity expansion cycles, where excess capacities of both attributes are required to reach a desired minimum level at the start and the end of each cycle. Following Microsoft's demand trends, we specifically focus on exponential demand growth. Using piece-wise linear demand approximation, we partially solve the problem, and to fully solve the problem, we devise a dynamic-programming-based algorithm. We next propose a forward-looking heuristic based on minimizing total cost rate in each cycle. We also examine the problem of cluster selection, given a set of available cluster-types. Last, we conduct a numerical study of the performance of the proposed heuristic.

Next, we focus on the effects 3D printing can have on manufacturer-retailer relation in supply chains. 3D printing is a relatively new manufacturing technology that is attracting attention from many firms and governments. However, its impact on operations and firms relationships in a supply chain remains unexplored. In the fourth chapter, we investigate the impact of 3D printing, or additive manufacturing, on a supply chain consisting of a manufacturer and a retailer. We study two scenarios: 1) 3D printers are utilized by the manufacturer; and in a more novel situation 2) 3D printing technology is adopted by the retailer. We analyze the equilibrium of Stackelberg games in both cases and compare the

results with a benchmark system without 3D printing. In the first scenario, both the retailer and the manufacturer are better off. In the second scenario, however, the retailer might be worse off although he is adopting 3D printing. This is because the supply chain is more responsive to demand, and the manufacturer can set higher wholesale prices for 3D plans and capture higher profits. We next compare the two scenarios. We identify and quantify the positive and negative effects associated with 3D printing for both firms in a supply chain. In many cases, the novel scenario of the retailer producing products, made possible by the unique features of 3D printing, results in the best profit outcomes for both firms.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to University of Washington, where I have had the opportunity to work with my advisers Professor Kamran Moeinzadeh and Pofessor Shi Chen. Prof. Moinzadeh's interest in Inventory Management inspired me to work in this area, and together with Prof. Chen and Prof. Wagner (my coauthor), they have immensely helped me in my research skills. I would also like to express my gratitude to Prof. Moinzadeh and Prof. Studer-Ellis to have guided me to improve my teaching skills.

Next, I am deeply thankful to my family who have been substantially patient and supportive of my PhD. Due to the current issues between Iran's government and the United State's, neither can I travel to Iran, my home country, to visit my family, nor can they travel to the US. As the result, I have not seen my family for 5 years. Although this can be disappointing sometimes, my family has always encouraged me to pursue my PhD.

Last but not the least, I am grateful to my fellow Foster PhD students for sleepless nights we had in Mackenzie Hall and all the fun in the past five years. I am so thankful to Shahryar Doosti, Pegah Jalili, Behnaz Gh, Sareh Nabipour, Amir Fazli, Soraya Fatehi, Eugene Pavlov, Aravinda Garimella, Emisa Nategh, Majid Majzoobi and Ebrahim Barzegar.

DEDICATION

To my beloved parents,

Reza Arbabian and Sakineh Esbati

Chapter 1
INTRODUCTION

Emerging technologies are defined as technologies whose developmental and/or practical applications are still largely unrealized. Rotolo el al (2015) propose a framework, which based on: 1) radical novelty, 2) relatively fast growth, 3) coherence, 4) prominent impact, and 5) uncertainty and ambiguity, classifies a technology as emergent. Using their framework, Weldon (2019)[1] studies different technologies, and he ranks the first ten emerging technologies in 2019 as follows: 1) computer vision, 2) deep learning, 3) natural language processing, 4) business-wide networking fabrics, 5) edge computing, 6) quantum computing, 7) Nano technology, 8) cloud computing, 9) additive manufacturing and 10) augmented reality. While adopting each of these technologies can substantially affect operations of a firm, we specifically focus on cloud computing and additive manufacturing, because they pose new challenges for matching supply and demand in a firm or in a supply chain.

Matching supply and demand has long been studied in the operations management field. The introduction of the book *Matching Supply with Demand* by Cachon and Terwiesch (2008) discusses the importance of this topic in detail. A common objective in my dissertation is to study how firms or supply chains can design their operations to better match supply and demand when new technologies emerge. In specific, I focus on cloud computing and additive manufacturing. Consider the following examples:

- Users attempting to access T2 micro instances (which are a low-cost, general purpose

[1]https://www.information-management.com/slideshow/12-top-emerging-technologies-that-will-impact-organizations

virtual servers that provide a baseline level of CPU performance with the ability to burst above the baseline when needed) in Asia Pacific North East region in August 2014 were told capacity had maxed out, barely after release in July.

- In 2017, Amazon's brand-new UK T2 micro instances reached saturation point, with users being told the AWS service had run out of local capacity.

- In 2017, Ministry of Supply, which is a garment company, installed a 3D printer in one of its retail stores in Boston. This retailer can 3D print different types of garments in less than 90 minutes and deliver it to the customers in the store.

The first two examples have in common that they suffer from a mismatch between demand and supply. The third example, however, attempts to decrease supply and demand mismatches via unique features of 3D printing.

Regarding cloud computing, numerous references (e.g., Gartner Press Release 2017 and Cisco 2018[2]) report and predict a surge in demand for cloud services in recent years and in future. For instance, Figure 1.1 shows global cloud traffic has grown from 3.4 to 10.4 zettabytes per year since 2014 with compound annual growth rate of 25% (Cisco 2016).

Demand for cloud services is satisfied by four major attributes: 1) storage, 2) network, 3) central processing unit (CPU), and 4) random access memory (RAM). All of these attributes agglomerated in data centers are the hardware infrastructure of all cloud services. As demand grows over time, these attributes can potentially result in a capacity bottleneck. In case of a shortage in storage or network, the capacity of these two attributes can be scaled up individually. However, if the capacity of RAM or CPU becomes the bottleneck, the company replenishes new servers.

---

[2]https://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.html

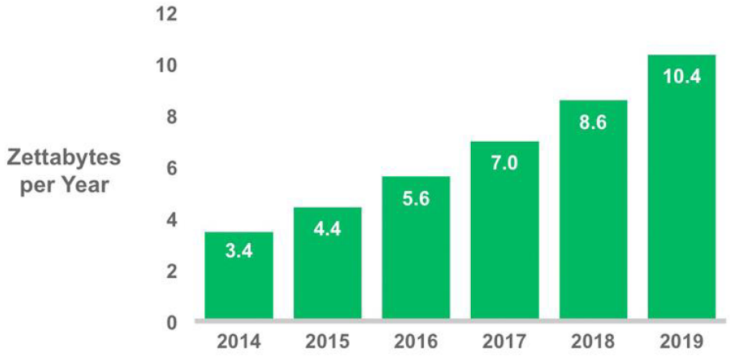


Figure 1.1: Global cloud traffic growth.

One distinctive feature of this industry is that each server increases the capacity of RAM and CPU simultaneously. That is, depending on the server configuration, one unit of supply (i.e., server) increases a fixed ratio of CPU to RAM. Therefore, on the demand side, the growth of CPU and RAM is time dependent and predictable over time (Shen el al. 2014), and on the supply side, the increase in the capacity of RAM and CPU takes place simultaneously via replenishment of pre-configured servers. Since the demand ratio is disproportionate to the supply ratio, imbalances between supply and demand can increase. To match supply and demand, more efficiently, in a cloud company, we focus on a capacity-expansion model where the provider employs two different types of supply packages (i.e., servers). We further discuss under what economic conditions employing only two supply packages is no different than employing many. Therefore, the unique feature of expanding the capacity of RAM and CPU in cloud industry is: demand growth rates for the two attributes are time-dependent, whereas the addition of these two attributes is "bundled" in pre-configured supply packages.

Next, in Chapter 4, we focus on additive manufacturing, also known as 3D printing, in which many governments (e.g., USA, China, India, Japan, France and Germany) as well as industry leaders (e.g., GA, BMW, Disney and Mercedes Benz) are investing. According to

Figure 1.2[3], the first four reasons why companies invest in additive manufacturing are: 1) prototyping, 2) product development, 3) innovation and 4) increasing supply chain efficiency. In Chapter 4, we investigate the fourth reason.

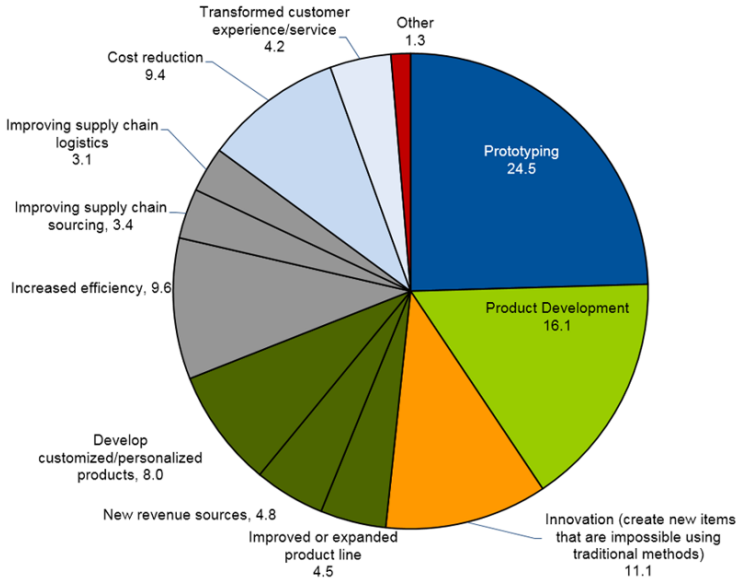


Figure 1.2: Why companies invest in 3D printing?

To study how 3D printing affects supply chains, one should be aware of: 1) how it is operated, 2) what are the costs associated with the technology, and 3) what are its benefits. 3D printing a product begins with a 3D plan model which is sliced into fine layers. These layers are, then, turned into instructions for a 3D printer by a computer. Eventually, the 3D printer deposits material layer by layer till the model is replicated (Hannon (2019)). Therefore, there are 3 costs associated with adopting 3D printing technology: 1) cost of purchasing 3D printers, 2) cost of developing/purchasing 3D plans, which is a fixed cost and is independent of the number of products being 3D printed, and 3) cost of material/filament, which is a linear function of the number of products being 3D printed. On the other hand,

[3]https://www.fabbaloo.com/blog/2014/12/11/gartner-identifies-the-big-barrier-to-3d-printing

the benefits of additive manufacturing are: 1) 3D printing an object with a complex shape has the same cost as 3D printing an object with a simple shape, and 2) 3D printers, in supply chains, can be utilized close to final customer (e.g., they can be utilized in a retail store) to satisfy the demand more responsively.

To investigate these effects, we focus on a simple supply chain consisting of a manufacturer and a retailer serving stochastic customers, and we study the following two questions: 1) under what economic conditions should the manufacturer adopt 3D printing, and 2) under what economic conditions should the retailer adopt 3D printing. An obvious benefit of retailer 3D printing is the elimination of transportation costs of completed products; however, we will identify additional benefits of this unique setup (e.g., higher supply chain products). Since it is uncommon for a retailer to use traditional manufacturing techniques, we see the possibility of a retailer 3D printing products as a unique feature. That is, the whole supply chain can be potentially more responsive to demand.

1.1 Scope and Contributions

In Chapter 3, after studying the literature in Chapter 2, we focus on a capacity-expansion model of two attributes (i.e., RAM and CPU) of a cloud infrastructure provider. This problem has a distinct feature: demand growth rates for the two attributes are *time-dependent*, whereas the addition of these two attributes is "bundled" in pre-configured supply packages. Employing only one type of pre-configured supply package will create an imbalance between the supply and the demand of the attributes. Thus, the provider needs to employ at least two different types of supply packages to achieve the supply-demand balance, such as a CPU-intense package and a RAM-intense package. In theory, the provider can design and employ many types of server configurations. However, there is a preference for the minimum variety of server configurations in the data centers for the sake of maximum fungibility and

productivity. Therefore, this chapter focuses on capacity expansions with two types of pre-configured supply packages. We also will examine conditions under which the employment of two types of supply packages can be as good as the employment of many types. This chapter makes three major contributions to the literature.

1. First, to our knowledge, the problem considered in this chapter, to which we refer as a *capacity expansion with bundled supplies of capacity attributes problem*, has not been studied. This problem is important in practice and has major implications in a rapidly growing industry. According to Frye (2013), approximately 5.75 million new servers are purchased by cloud providers every year; however, about 30%, after installation, are left unused for six months or more (Kepes 2015). These numbers indicate that the capacity-expansion policies of the cloud infrastructure providers, which are seen mainly through the acquisition of servers, are crucially important and have serious cost implications. The expansion costs are not due merely to the replenishment and installation of servers but also to the significant financial costs associated with having excess capacities, a well-recognized problem in the capital-intensive cloud industry. Hence, we believe that the problem studied in this chapter is timely and important.

2. Second, the capacity expansion with bundled supplies of capacity attributes problem is different from classical single-resource capacity-expansion problems. The key difference is that, in this problem, one unit of supply increases the capacities of two attributes simultaneously, but the ratio of the supply of the two attributes does not necessarily match the ratio of their demand growth rates. In such a problem, the concerns that need to be investigated are: (i) the optimal determination of the expansion times and quantities of the two types of supply packages; and (ii) the optimal selection of the configurations of the two supply packages from a variety of possible configurations.

Moreover, an important issue to address is when the employment of only two types of supply packages can be good enough. Our study aims to address all of these problems.

3. Third, we analyze capacity-expansion policies for a finite planning horizon, which are motivated by two common policies that we have observed in practice. The first policy is to employ mainly one type of supply package in consecutive expansions throughout the planning horizon. Such a policy initially results in the accumulation of excess capacities of one of the attributes, which can be dealt with by adding a single expansion with a different type of supply package toward the end of the horizon. For example, the provider can employ a CPU-intense type of package in consecutive expansions, followed by a final expansion by employing a RAM-intense type of package such that demand and supply of the capacities of both attributes are leveled at the end of the planning horizon. In contrast, the second policy is to employ a mixture of both types of supply packages at every expansion point such that excess capacities of both attributes are leveled right before the next expansion point. We refer to the first as the *sequential-replenishment policy* and the second as the *simultaneous-replenishment policy*. We will show that, once optimized, whether a policy dominates the other depends on the potential savings on the fixed expansion costs due to the joint replenishments of the simultaneous-replenishment policy.

In the Chapter 4, we propose and analyze stylized game-theoretic models of a simple manufacturer-retailer supply chain for a single product, where 3D printing is available. We first consider the natural case where the manufacturer may purchase one or more 3D printers to supplement/replace traditional manufacturing techniques. One example of such a manufacturer is General Electric, which produces fuel nozzles tip for LEAP engine, using 3D printing. In this case, on the one hand, all three costs associated with 3D printing are

incurred at the manufacturer. On the other hand, the manufacturer benefits from the fact that 3D printing objects with complex shapes is cheaper than producing products using traditional manufacturing techniques. Furthermore, we introduce a novel arrangement where the *retailer* may utilize 3D printers; an example of such a retailer is Ministry of Supply in Boston, MA (Schiffer 2017). In this case, the cost of purchasing 3D printers and cost of filament are incured at the retailer while the manufacturer bears the cost of developing 3D plans. An obvious benefit of retailer 3D printing is the elimination of transportation costs of completed products; however, we will identify additional benefits of this unique setup (e.g., higher supply chain profits). Since it is uncommon for a retailer to use traditional manufacturing techniques, we see the possibility of a retailer 3D printing products as a unique feature. We utilize a natural benchmark to evaluate the impact of 3D printing on the supply chain: a decentralized system with no 3D printing available. Next, if 3D printing is adopted by either the manufacturer or the retailer, it is crucial to note that the capacity of one 3D printer is limited. Therefore, the firm should decide about total 3D printing capacity, which is equivalant to opting for the number of 3D printers. Eventually, given the available 3D printing capacity, the firm should decide how many of the products to 3D print. The following is a summary of this chapter's main contributions to the operations management literature.

1. The unique characteristics of 3D printing result in its potential adoption by either the manufacturer or retailer, resulting in new models of cash, material and information flows.

2. For the case where the manufacturer may adopt 3D printing, we analytically derive the equilibrium of the interaction between manufacturer and retailer. We show that a necessary condition for the manufacturer to adopt 3D printing is that the unit 3D

printing cost must be at most the unit cost of traditional manufacturing. Furthermore, the fixed cost of developing 3D printing plans from the traditional manufacturing plans must not be too costly. Production will be either 3D printed or produced traditionally, with no hybrid solution possible. Finally, if 3D printing is economically feasible for the manufacturer to utilize, then it benefits both the manufacturer and retailer, with respect to the benchmark system.

3. For the case where the retailer may adopt 3D printing, we analytically solve the retailer's subproblem characterizing his behavior. We can then partially solve for the equilibrium under the additional assumption of uniformly distributed demand; accompanying numerical results, including for a normal distribution of demand, indicate that our analytical results closely mirror the actual equilibrium. Interestingly, 3D printing may be adopted even if the unit printing cost is more than that of traditional manufacturing; this is due to the increased responsiveness of the 3D printers at the retailer location, allowing a make-to-order strategy that is not possible at the manufacturer. We show that there are many economic and competitive scenarios where the retailer's adoption of 3D printing leads to improved profit outcomes for both the retailer and manufacturer, with respect to the benchmark system. Finally, in contrast to the case where the manufacturer adopts 3D printing, it is possible to have a hybrid equilibrium where traditional manufacturing and 3D printing are utilized in tandem.

4. Comparing the scenario where the manufacturer adopts 3D printing to that where the retailer adopts 3D printing, one learns that the manufacturer and the entire supply chain always prefer the *retailer* adopting 3D printing. The retailer's preference, however, depends on problem parameters, and can prefer either manufacturer or retailer adoption of 3D printing. Notably, there exist supply chain parameters where both

the manufacturer and retailer (and consequently the supply chain) prefer the retailer's adoption of 3D printing, underscoring the importance of this novel combination of a retailer producing products using a new manufacturing technology.

In Chapter 5, we summarize our findings and provide research extensions. For all the proofs in this dissertation, we refer the reader to the Appendix.

Chapter 2

LITERATURE REVIEW

In Sections 2.1 and 2.2, we study the literature relate to the cloud and 3D printing problem introduced in the Introduction.

2.1 Cloud Computing

Our study of cloud computing is related to three bodies of the literature: capacity management and expansion, multi-item inventory systems, and operational strategies for the cloud computing industry.

2.1.1 Capacity Expansion

The literature on capacity management and expansion is vast and dates back to the classic paper by Manne (1961). He explains that the trade-off in capacity expansion problems is between the economies of scale and an anticipated persistent growth in demand for capacity, as shown in Figure 2.1. He finds the optimal time and quantities of the expansions. He assumes the installation costs that result from a single capacity increment of size x has the following parametric function:

$$kx^a \quad (k > 0; 0 < a < 1)$$

where a and k are constant values. He next adopts the expression e^{-rt} as the present value of a dollar due in t years in the future. Therefore, the total cost of the capacity expansion

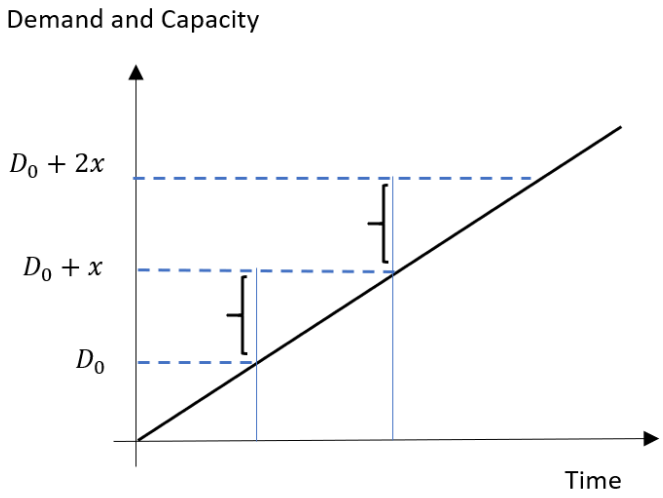


Figure 2.1: Growth of demand and of capacity over time.

problem modeled in his paper is as follows:

$$C(x) = kx^a + e^{-rx}C(x)$$

where C(x) is a function of x that represents the sum of all discounted future costs. Next, the optimal order quantity is the unique solution of the following equation: $a = \frac{r\hat{x}}{e^{r\hat{x}}-1}$, where $\hat{x}$ is the optimal order quantity. He then models stochastic demand and shows the optimal order quantity under stochastic demand has similar expression to deterministic demand.

Similar to Manne's study, the trade-off in our cloud computing problem is between the economies of scale and an anticipated persistent growth in demand for capacity. However, in our problem, one unit of supply increases the capacity of multiple attributes simultaneously.

Subsequently, Luss (1982) and Martinez-Costa (2014) reviewed the studies on capacity management. These reviews found that most capacity expansion problems are concerned with sizes, times, and locations (or capacity types) of expansions with a trade-off between

the holding costs of excess capacities and economies-of-scale savings (e.g., fixed expansion costs independent of expansion sizes). One stream of work in this area, as related to our problem, provides capacity expansion models for multiple facilities or multiple capacity types. Typically, multi-facility problems assume that the same capacity type is added to different locations and that there is an additional shipping cost when demand at a location is satisfied using the capacity of a different location (e.g., Fong and Rao 1975). Erlenkotter (1974) considered a multi-facility model whereby there are different capacity types, namely, the deluxe capacity and the standard capacity. Similar to Erlenkotter (1974), Eppen et al. (1989) and Li and Tirupati (1994) considered a multi-facility model in which capacity comes in the forms of flexible capacity and the dedicated capacity. Note that the standard/dedicated capacity can satisfy only one particular type of demand, whereas the deluxe/flexible capacity can satisfy either type of demand. Such a model is similar to that of our problem, as there are two streams of demands and two types of capacities present.

There is a fundamental difference between the above-noted models and ours. In those models, a unit of supply provides a unit of capacity, which can be used to satisfy only a particular type of demand. The deluxe/flexible capacity can be converted to meet either type of demand, but it cannot provide capacities to meet both types of demands simultaneously. In contrast, in our work, two types of capacities (i.e., two attributes) are both included in a package, and thus a unit of supply provides capacities to simultaneously satisfy both demand types. Moreover, the ratio of the attributes included in a package is different across different types of packages.

2.1.2 Inventory Management

Our study is also related to the inventory management, including the multi-item inventory systems, literature. Using the terminology from inventory theory, we define demand as,

indeed, the incremental demand for new capacity and replenishment as the new capacity added in an expansion. Thus, immediately after an expansion, there would be excess capacity, which would be depleted as demand for new capacity occurs. Hence, the capacity-expansion problem can be analyzed in the same way as an inventory problem.

In the inventory management literature, the studies of dynamic economic lot-size (DEL) models are related to this study. The DEL models can be viewed as periodic review models with fixed ordering and holding costs. Demand, as well as costs, can vary with time. The DEL literature dates back to the seminal work by Wagner and Whitin (1958), who proposed a forward dynamic programming procedure for solving multi-period lot-sizing problems. Later studies focused on the design of heuristics to determine lot sizes based mainly on single-product models, such as the well-known study by Silver and Meal (1973). Baker (1989) provided a comparison of various heuristics. More recently, studies have focused on rolling schedules and the impact of a truncated planning horizon, e.g., Stadtler (2000), Fisher et al. (2001), and Van Den Heuvel and Wagelmans (2005).

Another body of inventory management literature that is related to cloud problem is the joint replenishment problem (JRP). The JRP is considered difficult to solve (especially when demand is time-varying), and numerous heuristics have been proposed, e.g., Silver (1976), Goyal (1974). See also Khouja and Goyal (2008) for a review of the heuristics developed for JRP. It is important to note that, typically, JRP allows arbitrary quantities of items in an order or package, whereas our problem assumes that the quantities of items (i.e., attributes) in a package are pre-determined. Thus, the challenge in our setting arises mainly from the intricacies of using the pre-configured supply packages to satisfy the time-varying demands for different attributes.

There are a few studies, such as those of Luss (1984), Lu and Qi (2011), and Agrawal and Smith (2017), that consider supply in packages with a fixed ratio of multiple products.

Agrawal and Smith (2017)'s paper is perhaps the most recent relevant study to ours. They focus on stochastic demand, and the time horizon, in their study, is assumed to be finite. Moreover, one unit of supply increases a fixed ration of multiple SKUs simultaneously. They define:

- t: number of time periods to go before the end of the season,
- ξ_{it}: shipping and handling cost for sending one prepack of size i through the supply chain to be ready for sale at the store at time t.

Next, the inventory shipped to the store in period t is specified by the following variables:

- X_{it}: the number of prepacks of size i that are shipped to the store in period t,
- s_t: the total amount shipped in period. Therefore,
- $s_t = \sum_i iX_{it}$ with a corresponding cost $\sum_i \xi_{it}X_{it}$

The optimal s_t is determined at the beginning of period t based on the ending inventory in period $t+1$. The other variables for the newsvendor model at the store are as follows:

- E_t: the inventory at the end of period t.
- $I_t = E_{t+1} + s_t$: the beginning inventory after ordering in period t.
- $p_t(k) = P\,(\text{demand} = k \text{ in period } t)$.
- $P_t(k) = P\,(\text{demand} \leq k \text{ in period } t)$.
- μ_t: mean demand in period t.

- π_t: retail price per unit in period t.
- c: retailer's unit cost of the item (in addition to costs captured by ξ_{it}).
- c_{lt}: additional loss per unit short (in addition to $\pi_t - c$), such as loss of goodwill.
- c_s: salvage value per unit at the end of the last period.
- h_t: holding cost per unit.

For the case of lost sales, the ending inventory for time period t is equal to one of the discrete values $E_t = 0, 1, ..., I_t$. The probability distribution for the ending inventory conditional on I_t can therefore be written as

$$\Phi_t(E_t|I_t) = \begin{cases} 1 - P_t(I_t - 1) & E_t = 0 \\ p_t(I_t - E_t) & 0 < E_t \leq I_t \\ 0 & E_t > I_t \end{cases} \tag{2.1}$$

The expected newsvendor cost can be written:

$$C_t(I_t) = (c_{ot} + c_{ut}) \sum_{k=0}^{I_t - 1} P_t(k) + c_{ut}(\mu_t - I_t) \text{ for } t = 1, 2, ..., T.$$

where $c_{ut} = \pi_t - c + c_{lt}$, and $c_{o1} = c - c_s$ for the last period, and $c_{ot} = h_t$ for all $t > 1$. Therefore, newsvendor's profit function at time t is:

$$\Pi_t(I_t) = (\pi_t - c)\mu_t - C_t(I_t).$$

Note that prepack of size i may or may not be part of the set of prepacks in use. Thus, they define y_i a binary decision variable representing whether or not prepack of size i is ordered.

Next, for a given set of prepack types, they define a binary vector of $y = (y_1, y_2, ..., y_P)$. Then, the total expected profit for time period t through 0 is defined as $W_t(E_{t+1}|y)$= expected profit for the remaining t periods, given that the ending inventory from the previous preiod is E_{t+1}. with t periods remaining, the DP recursion relationship is

$$W_t(E_{t+1}|y) = \max_{X_{it}} \left[\Pi_t(I_t) - \sum_i \xi_{it} X_{it} + \sum_{E_t=0}^{I_t} \Phi_t(E_t|I_t) W_{t-1}(E_t|y) \right]$$

where, $I_t = E_{t+1} + \sum_i iX_{it}$, for $t = 1, 2, ..., T$, and X_{it}'s are integer, and $W_0(E_1|y) = 0$ for all E_1. The probability distribution of the initial inventory E_{t+1} is assumed to be known (e.g., it could be 0 with probability 1). They next argue that this problem is difficult to solve because of the integer decision variables and the fact that objective function is non-linear. Then, they simplify the problem to the case in which the shippinig and handling costs are independent of t. With this assumption, they pre-calculate the optimal number of prepacks size i, and use it in the DP. Then, in the steady state, they simplify the problem into a linear programming formulation, which is easily solvable.

The challenges addressed in Aggrawal and Smith (2018) and the other literature considering supply in packages with a fixed ratio of multiple products are similar to those addressed in ours. That is, how to "take care of the heterogeneous demands of multiple products" using the pre-configured packages (Lu and Qi 2011). Nevertheless, Luss (1984) and Lu and Qi (2011) studied the use of only one type of pre-configured package, and they allowed shortages to balance capacities, making their studies different from ours. Agrawal and Smith's (2017) study is probably the most related study to our work, as they focused on the optimal replenishment policies for an inventory system, whereby multiple SKUs are shipped together as "pre-packs," and similar to our work, the authors also discussed the optimal configuration of the pre-packs. They assumed, however, that demands for SKUs are stochastic and

stationary, whereas we consider that demands are deterministic and non-stationary.

Another stream of literature which resembles our study of cloud industry is co-production. Most studies in this literature consider only one type of input (or technology) that yields multiple types of final products, and focus on production quantity, product pricing and inventory allocation decisions for such co-production systems (Nahmias and Moinzadeh (1997), and Tomlin and Wang (2008)). There are a few exceptions that consider two or multiple types of inputs with different co-product splits. Motivated by the beef processing industry, Boyabatl et al (2011) analyzed a firm's optimal procurement quantities from two sources: contract and spot purchase, and they focused on the impact of the long-term contracts and spot market characteristics on the optimal decisions. Ng et al (2012) considered an arbitrary number of types of inputs and developed a robust-optimization model to efficiently determine a co-production firm's production and inventory allocation decisions. Dong et al (2014) analyzed a refinerys optimal portfolio of inputs (i.e., the refinerys processing range flexibility), and examined the impact of the portfolio on the value of downward substitution of the final products (i.e., the refinerys conversion flexibility). Chen et al (2017) considered two co-production technologies, which dier in their processing costs, overall yields, and co-product splits. They not only derived the optimal production and pricing decisions for a monopoly co-production firm, but also examined the impact of the characteristics of the two technologies on the firms optimal technology adoption decision.

Although related, our study of cloud industry has major differences from the existing literature on co-production. That is, we consider a firms optimal procurement of two types of supply packages (inputs) faced by demands for two non-substitutable products, with a focus on deriving the optimal inventory (excess capacity) control policies in a multi-period, finite-horizon setting, whereby the demand growth rates of the two products are time-varying and disproportionate.

Even though there is a relatively small body of work in which supply is a bundled collection of items, there is a rich body of literature on models of "bundled demand" for multiple items, especially in the area of assemble-to-order (ATO) systems. ATO systems are production-inventory networks, where multiple components are to be kept in stock and used to assemble orders that consist of multiple products. Due to the correlation across different products in demand packages, analyzing the inventory control and allocation policies of the ATO systems has attracted much attention. Song (1998, 2002) derived the operating characteristics of an ATO system, and Lu et al. (2003, 2010) analyzed alternative inventory allocation policies for ATO systems. Song and Zipkin (2003) provided a comprehensive review of ATO systems.

2.1.3 Strategies for Cloud Industry

In regard to the operational strategies of cloud computing providers, in the last decade, where the research interest has been growing. One stream of studies focuses on the pricing strategies of the cloud providers who have a large population of customers. Chen et al. (2018a) analyzed two different pricing schemes, reservation-based and utilization-based, employed by Amazon Web Services (AWS) and Google Cloud Platform (GCP), respectively, for their committed services. Li and Kumar (2018) examined an incumbent cloud provider's subscription-based pricing scheme as faced by an entrant provider. There also are a number of studies that focus on pricing schemes for the cloud providers' preemptible or low-priority services, such as the spot-price scheme of AWS (Xu and Li 2013, Kilcioglu and Maglaras 2015, Toosi et al. 2016) as well as the uniform-discount scheme of GCP and Microsoft Azure (Chen et al. 2018b).

There also is a body of literature on cloud capacity management. Some studies focus on the capacity allocation problem as faced by uncertain demand arrivals, which come with

various computing requirements (Dieker et al. 2016) and have considered the possibility of migrating queries between machines to optimize capacity utilization (Van et al. 2009). Other studies consider the provider's decision about the number of instances to deploy, with the associated deployment costs, namely, the elasticity management problem. The objective is typically to minimize the total instance deployment cost subject to a service-level-agreement constraint. See Galante and Bona (2012) for a comprehensive review of the literature on cloud elasticity management.

While the literature on cloud computing capacity management has focused on short-term (e.g., hourly, daily) decisions, including instance deployment, allocation, and reassignment, our study focuses on mid to long-term (e.g., monthly, quarterly) decisions about capacity expansions. In practice, the cloud firm calls on its forecasting team to provide a projection of the demand growth for a finite planning horizon (e.g., six months or one year). Then, using the projected demand growth as an input, our model can be applied to the cloud firm's capacity-expansion decisions.

2.2 Additive Manufacturing

Our study of adopting 3D printing technology in a supply chain is related to three bodies of the literature: dual sourcing, supply chain contracting, and operations of 3D printing.

2.2.1 Dual Sourcing

In our study, the scenario where retailer may adopt 3D printing resembles dual sourcing (e.g., Veeraraghavan and Scheller-Wolf (2008)). Ramasesh et al (1991) study simultaneous procurement from two sources when supply has uncertain lead time. In their model, demand is constant and they focus on reorder point, order quantity inventory models. Yu el al (2009) study the impacts of supply disruption risks on the choice between the single and dual sourc-

ing methods in a two-stage supply chain with a non-stationary and price-sensitive demand. Wang et al (2010) characterize the optimal procurement quantities and improvement efforts for the case where a firm can source from multiple suppliers and/or exert effort to improve supplier reliability, and they study both random capacity and random yield types of supply uncertainty. Chen and Yang (2014) consider a supply chain similar to what we study, and they allow the retailer to satisfy his shortages from an emergency back up supplier. Janakiraman and S. Seshadri (2017) study inventory systems with backordering under presence of dual sourcing.

Perhaps the most recent relevant study to ours is Arikan and Jammerneg (2014). Their setting is as follows: the firm can order the products from an off-shore supplier with long lead time, or from and on-shore supplier with short lead time. However, the latter incurs an extra cost. They define:

- $f(\cdot)$: demand distribution,
- q: order quantity,
- p: selling price,
- c: each unit's cost,
- d: extra cost of ordering from the on-shore supplier.
- s: salvage value,
- $L(q)$: expected amount of leftovers. Therefore, $L(q) = \int_0^q (q-x)f(x)dx$,
- $Q(q)$: expected lost sales. Therefore, $Q(q) = \int_q^\infty (x-q)f(x)dx$.

Therefore, if all the products are being ordered from the off-shore supplier, the firm's objective function is:

$$\pi(q) = (p - c)q - (p - s)L(q).$$

The solution to the above problem is Newsvendor's solution, i.e., $F^{-1}(\frac{p-c}{p-s})$. However, if the firm can order both from the off-shore and on-shore supplier, then her objective function is:

$$\pi(q) = (p - c)q - (p - s)L(q) + (p - c - d)Q(q).$$

where q is the number of products being ordered from the off-shore supplier. Therefore, the optimal order quantity is $F^{-1}(\frac{d}{d+c-s})$.

We model the adoption of 3D printing at the retailer's level similar to Arikan and Jammerneg (2014) and dual sourcing literature. However, the main difference between our study and the mentioned literature is that: 1) in dual sourcing one basic assumption is that the source with shorter lead time is strictly preferred to a source with longer lead time when the latter is costlier than the former. In our study, however, we do not have this assumption, and the retailer might opt for 3D printing whether 3D printing is costlier than traditional manufacturing or not. 2) Our model has a different cost structure. That is, 3D printing replenishment cost is incurred at the retailer, and 3D printing capacity is, also, decided by the retailer. 3) In dual sourcing literature, wholesale prices are exogenously set, while in our study, the manufacturer is the leader and sets wholesale prices.

2.2.2 Supply Chain Contracting

In the past few decades, an extensive literature has been developed that focuses on supply chain contracting. For an overview of this literature, we refer the reader to Cachon (2003) and Hohn (2010). An important goal in many of these studies is supply chain efficiency,

where the double marginalization effect is reduced. In an early study, Pasternack (1985) shows that a buy-back contract can coordinate the supply chain. Cachon and Lariviere (2005) show that a revenue-sharing contract can also coordinate a supply chain, and they show it is theoretically equivalent to a buy-back contract. Tsay (1999) shows that quantity-flexibility contracts can also coordinate a supply chain. However, Wang andHu (2010) show that when demand is advertisement dependent, buy-back, revenue-sharing, and quantity-flexibility contracts do not necessarily lead to coordination. Taylor (2002) shows that coordination is possible under sales-rebate contracts. In our study we utilize the simple wholesale-price contract and variants thereof; however, the differentiating feature in our modeling is the presence of a 3D printer.

More broadly, introducing 3D printing into a supply chain resembles investing in manufacturing to decrease the unit production cost or increase the production capacity. The problem of investing in production has long been studied in the literature. An early paper is Porteus (1985), which considers, in an EOQ setting, the set up cost as an endogenous factor that is a function of investment, and derives the optimal investment. In the flexible manufacturing literature, Mieghem (1998) extends the model in Fine and Freund (1990) and studies optimal investing in flexible manufacturing. Goyal and Netessine (2007) combines and studies the interaction of competition with technology investments. Our study also considers an investment in new technology, the 3D printer, by either the manufacturer or retailer; however, we show that higher 3D printing unit production costs can still be beneficial for the supply chain under the scenario where the retailer adopts the new technology.

Another relevant literature stream concerns cooperative research and development in a supply chain. Ge et al (2014) assume investment in production can take place at both the upstream supplier and the downstream manufacturer, and they derive the optimal investment strategies for both firms. Bernstein and kök (2009) consider a problem consisting

of a single assembler and multiple suppliers, and analyze the impact of different contracts on supplier investments. Ishii (2004) studies cooperative R&D from an economics perspective. Gupta (2008) studies how knowledge spillovers, resulting from manufacturer investment in process improvements, affect supply chain performance. Li et al (2012) show that a manufacturer sharing cost-reduction expenses with a supplier results in increased market share and higher profit for the supply chain. Our study also considers the cooperation between the manufacturer and retailer, albeit in a new and novel manner: the manufacturer incurs a fixed cost to develop 3D printing plans for possible use by the retailer, which pays for the right to use the plans. This unique arrangement can lead to improved profit outcomes for both firms.

2.2.3 Operations of 3D Printing

There is limited research in the operations management literature on the impact of 3D printing. We are aware of only a few papers that study different aspects of 3D printing. Song and Zhang (2016) study the tradeoff between 3D printing and traditional make-to-stock policies in spare-parts logistics, deriving optimal solutions in special cases which serve as high quality heuristics for a general model. Dong et al (2017) contrasts 3D printing with more traditional flexible manufacturing technologies, identifying the appropriate situations to apply 3D printing and the associated benefits. Both of these papers focus on the impact of 3D printing on a single firm; in contrast, we investigate the impact of 3D printing on a decentralized supply chain of two firms, a manufacturer and a retailer. Westerweel et al (2018) focus on the lower reliability of the 3D printed products. They compare 3D printing with traditional manufacturing methods using life cycle analysis. They further study the application of 3D printing in producing spare parts to reduce supply lead time and inventory cost. Sethuraman et al (2018) focus on the fact that consumers can 3D print, personal

fabrication. They characterize the market and conditions for personal fabrication. Chen et al (2017) study a, centralized, dual channel setting where the adoption of 3D printing is possible for each channel. In contrast to the above mentioned papers, we focus on a decentralized single channel supply chain, and study the conditions under which each player will adopt 3D printing.

The most recent relevant paper to our study is Chen el al (2018) where they consider two adoption cases of 3D printing in a dual-channel (i.e., online and in-store) retail setting, and evaluate its impact on a firm's product offering, prices for the two channels, as well as inventory decisions. They study the three cases shown in Figure 2.2. The three cases are as follows: 1) the base case in which the firm offers a finite number of product options, and uses BTS (build-to-stock) to satisfy in-store demands while BTO (build-to-order) for the online demands; 2) 3D printing is adopted at the factory so that it can use BTO and mass customization to satisfy online orders, while keeping BTS with a finite number of product options for in-store demands; and 3) 3D printing is also adopted at the stores, so that both online and in-store demands are satisfied under the BTO mode with mass customization options. They assume customers are heterogeneous in two dimensions: 1) the waiting cost

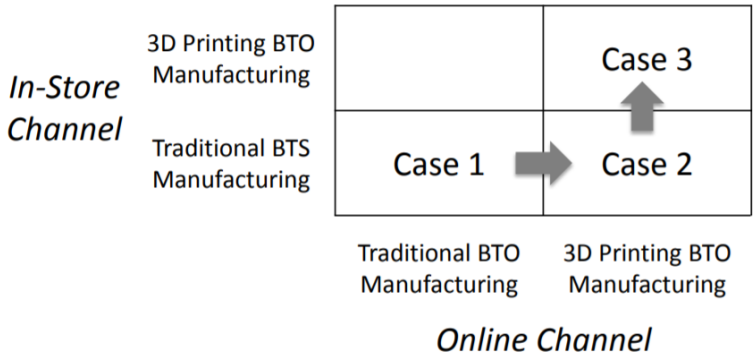


Figure 2.2: 3D printing technology adoption cases for consumer goods retailing.

they incur from purchasing online, and 2) the fit cost they incur because the product types offered by the firm do not meet their needs exactly. Specifically, customers' heterogeneous online waiting costs are captured by the three types of customers described as follows. Type I customers, which have zero online waiting cost, and they correspond to α propostion of the population ($0 < \alpha < 1$). Type II customers, which have positive online waiting cost, and they correspond to β proportion of the population ($0 < \beta < 1$ and $\alpha + \beta < 1$). Type III customers, which have infinite online waiting cost, and they correspond to $1 - \alpha - \beta$ proportion of the population.

They next model customers' heterogeneous product preferences using the circular city framework. They assume that customers are located on a circle of unit circumference. Customers are uniformly distributed on the circle. Each customer's location represents her ideal product type, and the arc distance between a product location and the customer location measures the customer's misfit from this product. Each customer only purchases the product type that is closest to her location on the circle. Given an in-store price p, the customer's utility from purchasing in-store a product that is x arc distance away is $v - p - tx$, where v is the valuation of customers for the ideal product type, t is the fit cost parameter and measures customers' sensitivity to product differences, and tx is the fit cost of customer x. Similarly, given an online price p, the customer's utility is $v - p - tx - e$ from purchasing online. Note that $e = 0$ for Type I customers and $e = \infty$ for Type III customers.

The total customer demand from the unit circle follows a normal distribution with mean μ and standard deviation σ. They assume that the demand at each point on the customer circle follows an i.i.d. normal distribution. Thus, the demand from a customer segment with arc length x follows a normal distribution with mean μx and standard deviation $\sigma\sqrt{x}$. Next, let $f_o(\cdot)$ and $f_i(\cdot)$ denote the normal probability density function for the online and in-store demands, respectively. Additionally, let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal

probability density function (pdf) and cumulative density function (cdf), respectively. And they assume that the firm incurs marginal cost $c = c_r + c_p$ for each unit of product, where c_r is the raw material's cost and c_p is the production cost. Next, when using the traditional technology, the firm incurs a production setup cost s for each product type it offers. Thus, given n product types, the total setup cost is sn. Finally, the firm incurs a fixed cost when adopting 3D printing. This fixed cost is k at the factory and k' at the store.

For each case of Figure 2.2, the firm needs to make four decisions: number of horizontally differentiated products to offer, prices for products sold online and in-store, as well as inventory decisions for the in-store channel. For the purpose of this section, we will mention only their study of Case 1. For the study of Case 2 and 3, we refer the reader to their paper.

In Case 1, the firm uses the traditional technology to produce n types of horizontally differentiated products, and chooses price p_o for all products sold online and p_i for all products sold in-store. Because product types are horizontally differentiated, the firm charges the same price for all product types within each channel. To derive the firm's optimal strategy, they need to first characterize the customer choices. Consider the arc on the customer circle that is centered at the location of any product type and has arc length $\frac{1}{n}$ $(n \geq 1)$. This arc corresponds to the demand base for this product type. Then, they derive the purchasing decisions of each type of customers as follows:

- Type I customers: Their utility from purchasing online is $v - p_o - tx$, and their utility from purchasing in-store is $v - p_i - tx$. Then, Type I customres purchase online if $v - p_o - tx \geq v - p_i - tx$ and $v - p_o - tx \geq 0$, purchase in-store if $v - p_o - tx < v - p_i - tx$ and $v - p_i - tx \geq 0$, and do not purchase otherwise. Thus, Type I customers purchase online if $p_i - p_o \geq 0$ and $0 \leq x \leq min\left(\frac{v-p_o}{t}, \frac{1}{2n}\right)$, and purchase in-store if $p_i - p_0 > 0$ and $0 \leq x \leq min\left(\frac{v-p_i}{t}, \frac{1}{2n}\right)$.

- Type II customers: Their utility from purchasing online is $v - p_o - tx - e$, and their utility from purchasing in-store is $v - p_i - tx$. Then, Type II customers purchase online if $v - p_o - tx - e \geq v - p_i - tx$ and $v - p_0 - tx - e \geq 0$, purchase in-store if $v - p_i - tx > v - p_o - tx - e$ and $v - p_i - tx \geq 0$, and do not purchase otherwise. Thus, Type II customers purchase online if $p_i - p_o \geq e$ and $0 \leq x \leq min(\frac{v-p_o-e}{t}, \frac{1}{2n})$, and purchase in-store if $p_i - p_o < e$ and $0 \leq x \leq min(\frac{v-p_i}{t}, \frac{1}{2n})$

- Type III customers: Their utility from purchasing in-store is $v - p_i - tx$, so Type III customers purchase in-store if $v - p_i - tx \geq 0$, and do not purchase otherwise. Thus, Type III customers purchase in-store if $0 \leq x \leq min(\frac{v-p_i}{t}, \frac{1}{2n})$.

Based on the customer choice, they obtain that the proportion of customers purchasing online is:

$$d_o(n, p_o, p_i) = \begin{cases} \left[\alpha min\left(\frac{v-p_o}{t}, \frac{1}{2n}\right) + \beta min\left(\frac{v-p_o-e}{t}, \frac{1}{2n}\right)\right] 2n & \text{if } p_i - p_o > e, \\ \alpha min\left(\frac{v-p_o}{t}, \frac{1}{2n}\right) 2n & \text{if } 0 \leq p_i - p_o \leq e, \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

and the proportion of customers purchasing each product type in-store is

$$d_i(n, p_o, p_i) = \begin{cases} 2(1-\alpha-\beta) min\left(\frac{v-p_i}{t}, \frac{1}{2n}\right) & \text{if } p_i - p_o > e, \\ 2(1-\alpha) min\left(\frac{v-p_i}{t}, \frac{1}{2n}\right) & \text{if } 0 \leq p_i - p_o \leq e, \\ 2min\left(\frac{v-p_i}{t}, \frac{1}{2n}\right) & \text{otherwise.} \end{cases} \tag{2.3}$$

Therefore, the firm's profit is:

$$\Pi(n, p_o, p_i, q) = \Pi_o(n, p_o, p_i) + \Pi_i(n, p_o, p_i, q) - ns,$$

where $\Pi_o(n, p_o, p_i)$ and $\Pi_i(n, p_o, p_i, q)$ are the profits from the online and in-store channels, respectively, and sn is the setup cost for n product types. Specifically,

$$\Pi_o(n, p_o, p_i) = (p_o - c)d_o(n, p_o, p_i)\mu$$

and

$$\Pi_i(n, p_o, p_i, q) = [p_i E\left(min(D_i, q)\right) - cq]\, n.$$

Using the Newsvendor's solution and the fact that demand follows a normal distribution, the last expression simplifies to:

$$\Pi_i(n, p_o, p_i) = \left[(p_i - c)d_i(n, p_o, d_i)\mu - p_i\phi(z^*(p_i))\sigma\sqrt{d_i(n, p_o, p_i)}\right] n$$

Therefore, firm's profit function simplifies to

$$\Pi(n, p_o, p_i) = (p_o - c)d_o(n, p_o, p_i)\mu + (p_i - c)d_i(n, p_o, p_i)\mu n - p_i\phi(z^*(p_i))\sigma n\sqrt{d_i(n, p_o, p_i)} - sn \tag{2.4}$$

Using 2.4, they derive the optimal decision variables for Case 1. They next analyze Cases 2 and 3, and through a set of numerical examples compare the 3 cases.

Similar to our problem, Chen et al (2018) focus on the adoption of 3D printing in a supply chain. They focus on a centralized supply chain where a firm serves customers in two channels. In our study, however, we focus on a decentralized supply chain with a singly channel, i.e., an off-line store.

Chapter 3

CAPACITY EXPANSION WITH A BUNDLED SUPPLY OF ATTRIBUTES: AN APPLICATION TO COULD COMPUTING

Over the past decade, technological enhancements led by the rise in cloud computing have revolutionized the IT industry. Cloud computing, as defined by National Institute of Standards and Technology, is "on-demand network access to a shared pool of configurable computing resources [that] can be rapidly provisioned and released with minimal management effort or service provider interaction" (Mell and Grance 2011). In recent years, we have witnessed a rapid growth in the cloud computing industry, represented by cloud service providers whose typical services are Software as a Service (SaaS) (e.g., salesforce.com), Platform as a Service (PaaS) (e.g., Google app engine), or Infrastructure as a Service (IaaS) (e.g., Microsoft Azure). For instance, according to Forbes, by 2014, more than half the businesses in the United States used cloud services[1]. Gartner, Inc. estimated that the worldwide public cloud services market grew 18% in 2017 to $246.8B compared to $209.2B in 2016 and that the IaaS segment was estimated to develop at the highest growth rate of 36.8% in 2017 to reach $34.6B (Gartner Press Release 2017). For cloud providers, such a huge surge in demand for cloud services has created a challenge in regard to how to plan their capacity expansions efficiently. This challenge could lead to lack of availability of cloud resources, which is the third reason, as shown on Figure 3.1, why companies hesitate in adopting cloud services, according to a survey by IDC in 2008 (Dillon et al. (2010)).

Capacity in cloud computing is usually defined by the availability of resources. Mishra et

[1]https://www.forbes.com/sites/reuvencohen/2013/04/16/the-cloud-hits-the-mainstream-more-than-half-of-u-s-businesses-now-use-cloud-computing

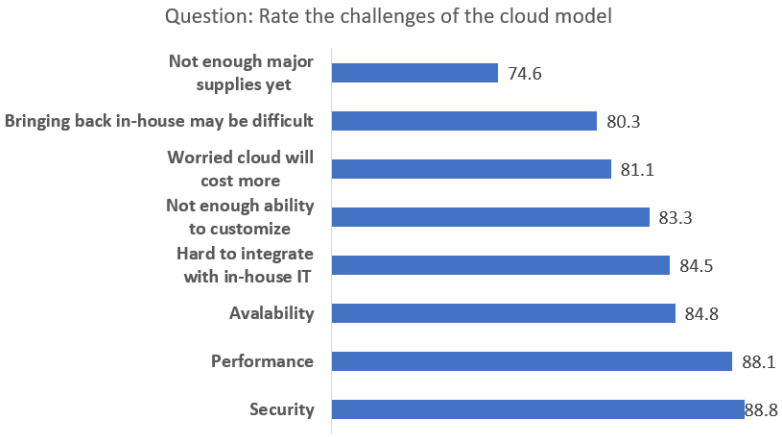


Figure 3.1: Cloud services adoption challenges

al. (2010) define the key resources in the cloud computing industry, referred to as capacity attributes in this chapter, as central processing unit (CPU), random access memory (RAM), network, and storage. These attributes account for the majority of spending in the industry. According to Dignan (2015): "76 percent of cloud spending is compute. . . . Network is 3 percent of cloud [spending] followed by 2 percent on storage and 4 percent [on] 'other'," where *compute* refers to CPU and RAM, the two main determinants of configurations of physical and virtual machines. All of these attributes agglomerated in data centers are the hardware infrastructure of cloud services. As demand grows over time, these attributes can potentially result in a capacity bottleneck. For example, to tackle this issue, Mishra et al (2010) propose a methodology for task classification in Google Cloud Backend.

In the case of a shortage in storage or network, the capacity of each of these two attributes can be scaled up individually. For example, cloud providers typically have storage units that are physically separate from the compute units. Thus, when there is a storage bottleneck, the capacity of storage can be added independent of other attributes. Similar issues that pertain to the network can be resolved either through an improvement in the network design or

by upgrading the network's individual components (e.g., switches and routers). Therefore, in the case of a shortage in these two attributes, capacity-related issues can be handled independently through the application of classic single-resource capacity-expansion models (e.g., Manne 1961).

In contrast to network and storage whose capacity can be scaled up individually, capacity expansion of the other two major resources, CPU and RAM, is managed through the acquisition of new servers. It is crucial to note that each server increases the capacity of RAM and CPU simultaneously; that is depending on the server configuration, one unit of supply increases a fixed ratio of CPU to RAM. Addressing the growing capacity requirements of CPU or RAM by the acquisition of the bundled supplies of those two attributes, instead of individual capacity scaling, occurs for two reasons: (i) cloud providers aggregate servers with similar configurations in clusters, and, as a result, the actual cost of individual capacity scaling, which includes not only the cost of purchasing individual CPU or RAM but also the cost associated with the clusters shut down during individual capacity scaling, is exorbitant; (ii) old server technologies may not be compatible with newer technologies, making the individual scaling of CPU and RAM cumbersome. (iii) The average life cycle of a cloud server is three years, and servers are commonly retired at the latest in five years. Therefore, when a server has operated for around four years, cloud companies would rather retire it than upgrade its RAM and/or CPU[2].

Capacity expansions of CPU or RAM through acquisition of the bundled supplies of those attributes, however, pose the challenge of achieving a supply-demand balance. On the demand side, even though the growing demand for CPU and RAM is relatively predictable over time (Shen et al. 2014), the growth rates of the two attributes can be *time-dependent* and *disproportionate*. On the supply side, the fixed ratio of CPU to RAM is not always

[2]https://www.pinterest.com/pin/478226054152078162/

equal to the time-varying ratio of the demand growth rates. Thus, to achieve supply-demand balance, a cloud provider should re-customize the configuration of the supply (i.e., servers) at any point in time. This, however, is not possible in practice for the reasons mentioned below. Therefore, inevitably, the provider would have to employ several types of supply configurations.

We conducted interviews with the individuals in charge of forecasting and capacity management for the cloud services of a leading public cloud infrastructure provider. We determined that the cloud provider often works with its hardware vendors to customize a few server configurations and then employs those pre-configured servers for an extended period of time, such as a planning horizon that typically spans six months to one year. The reasons for this process are as follows: (i) various engineering and technical constraints do not allow the choice of servers with an arbitrary ratio of CPU to RAM, limiting the choice of server configurations at the design stage; (ii) once the design of a new server configuration has been finalized, the provider commits to purchasing the customized servers over the lengths of the contracts with the vendors, which are usually several months to one year, and re-customization cannot occur too frequently as designing and testing a new configuration takes a significant amount of time (six months to one year) and is costly; and (iii) the company desires to reduce the variety of these customized servers in data centers, as increasing server fungibility can vastly improve cost efficiency through, for example, component and spares reduction, power reduction by server consolidation, and other supply chain-related efficiencies.

Although the motivation in this chapter is based on the cloud industry, the model and results are applicable to any other situations where one unit of supply feeds different units of products. One potential application is the chemical industry where chemicals with distinct grades are refined to different products. In this regard, an oil refinery industry is an apt

instance for the following reasons. 1) On the large scale, light and heavy crude oils are supplied to a factory. The two major products are gasoline and diesel fuel. From the former, however, more of each product can be refined, and as a result, it has a higher price. 2) Light crude oil is defined as having an American Petroleum Index (API) gravity between 32° and 42° by New York Mercantile Exchange, while heavy crude oil has an API less than 20°, meaning the former evaporates faster, and is much more toxic. 3) Demand rates for the final products (i.e., gasoline and diesel fuel) are not identical in different regions. For example, according to the US Energy Information Administration, gasoline and diesel fuel annual demands in 2016 were about 18 and 16 million barrels respectively in the US. For the above reasons, efficiently satisfying the demand for the two final products and managing the amount and timing of mixtures of the light and heavy crude oil supplies pose challenges similar to those of the capacity-expansion problem studied here. Thus, the results of our proposed optimal policy are applicable to the oil refinery industry.

3.1 Setting of the Base Model

We consider a capacity planning problem of a firm (e.g., a cloud provider) that faces demand for two attributes, $i \in \{1, 2\}$ (e.g., CPU and RAM), in a finite planning horizon (e.g., six months to one year). For example, in the cloud industry, a demand forecast for the purpose of capacity planning is given by the number of cores of CPU and the amount of memory. In this chapter, we focus on the firm's long-term capacity-expansion strategy when facing an increasing trend of demand for these attributes, but we do not consider the firm's capacity allocation decisions to tackle demand fluctuations on a day-to-day basis. Thus, we assume that the incremental (aggregate) demand rate for each attribute is non-negative but is time-dependent. For instance, demand for each attribute can exhibit an exponential growth with time.

In such a setting, capacities of the two attributes are added through replenishments of servers, which come in a variety of configured packages of the two attributes. In practice, servers with similar configurations are aggregated in clusters, and there are mainly two types of clusters: *CPU-intense* and *RAM-intense*. Hereafter, we use the terminology *cluster-type* j and *cluster* j interchangeably, where $j \in \{1, 2\}$, and note that a unit of a specific cluster-type contains a specific ratio of the two capacity attributes.

As noted, cloud providers such as AWS and Microsoft Azure work with their vendors on designing the cluster configurations, which takes significant time (six months to one year). If the planning horizon is not too long, the cloud providers employ only pre-configured cluster-types. Therefore, our base model focuses on the typical situation in which the cloud provider will use only pre-configured cluster-types throughout the planning horizon (Section 3.2), while we also extend the analysis to situations in which the cloud provider has a few opportunities to re-customize the cluster configurations within the planning horizon (Section 3.3.1). Let:

- T: length of the planning horizon.
- $d_i(t)$: incremental demand growth for attribute $i \in \{1, 2\}$ at time $t \in [0, T]$.
- $D_i(a, b)$: total demand growth for attribute i in time interval (a, b); $D_i(a, b) = \int_a^b d_i(t)dt$.
- a_{ij}: capacity units of attribute i in one unit of cluster-type $j \in \{1, 2\}$.
- c_j: unit replenishment (purchase) cost of cluster $j \in \{1, 2\}$.
- h_i: holding cost rate of excess capacity of attribute $i \in \{1, 2\}$.
- A: fixed expansion cost, independent of the size of expansion.

The holding cost includes the capital cost and depreciation of the excess capacity, while the fixed expansion cost includes the shipping, installation, and engineering costs of each expansion.

We assume that no capacity shortage is allowed. In other words, the cost of shortage is exorbitant. For instance, cloud providers typically aim at capacity levels that ensure an extremely high service level, defined as the percentage of time that demand can be satisfied immediately by available capacity. Leading cloud providers such as AWS, Microsoft Azure, and Google Cloud Platform have all promised higher-than-99% service levels in their service level agreements (SLA).

The assumption of no-shortage has another important implication for our base model. Admittedly, forecasts for demand growth throughout the planning horizon can never be accurate. In reality, the forecasting group may provide multiple projections of demand growth over time, namely, demand scenarios. Because the targeted service level is extremely high, however, the capacity planning group will often make expansion decisions based on a projected capacity requirement associated with the highest demand scenario. Therefore, in this chapter, we focus on the cloud provider's expansion decisions based on such a deterministic capacity requirement.

Our objective in regard to the problem studied here is to determine the times and quantities of the replenishments of two cluster-types, which minimize the total capacity expansion cost, that face a deterministic and time-dependent demand growth (capacity requirement) curve for each of the two attributes. The total cost consists of the fixed expansion costs, the holding costs of excess capacities of the two attributes, and the replenishment (purchase) costs of clusters, over the finite horizon T. Without loss of generality, we assume that capacities of both attributes are leveled, which means that the excess capacities reach a desired minimum level at the beginning and should be leveled again at the end of the planning hori-

zon. Moreover, we assume that capacity replenishment is immediate. Note that, when lead time for delivery is constant, it can be easily incorporated due to the deterministic capacity requirement curve for each attribute.

We focus here on a class of capacity expansion policies, for which the planning horizon is divided into multiple *capacity expansion cycles (CECs)*, and a CEC is defined as follows:

1. Capacities of the two attributes are added through replenishments of the two cluster-types such that capacities of both attributes are leveled at the start and at the end of the cycle.

2. Replenishments are made only when the capacity of at least one attribute is leveled.

3. Replenishments of the two clusters can be simultaneous or sequential.

As a special case, the planning horizon may consist of only one cycle. That is, capacities of the two attributes are required to be leveled again only at the end of the planning horizon.

The CEC policies have three important characteristics. First, replenishment of each cluster-type adds to the existing capacities of both attributes, although the ratio of the added capacities is fixed. For example, if the firm purchases Q_1 (or Q_2) unit of cluster-type 1 (cluster-type 2), this adds $a_{11}Q_1$ ($a_{12}Q_2$) of attribute 1's capacity and $a_{21}Q_1$ ($a_{22}Q_2$) of attribute 2's capacity to the existing capacities. Second, the planning horizon can be divided into multiple CECs with equal or different time intervals. That is, the number of CECs and their durations can be decision variables. Third, if the acquisition of the two cluster-types within a CEC is sequential, we consider the following policy in that cycle. The cycle may contain multiple purchases of a specific cluster-type, defined as the *leading cluster*, followed by a single purchase of the other cluster-type, defined as the *following cluster*, to level capacities of both attributes at the end of this cycle.

To summarize, in a finite horizon problem, a CEC policy contains N expansion cycles, where N can be a decision variable. In each expansion cycle $n \in \{1, 2, \ldots, N\}$, there can be multiple replenishment opportunities. If the replenishments of the two cluster-types are sequential, let m_n denote the number of replenishments of the leading cluster, followed by a single replenishment of the other cluster. If the replenishments of the two cluster-types are simultaneous, let $m_n + 1$ be the number of (joint) replenishments of the two cluster-types in cycle n. Note that the replenishment pattern may be different from cycle to cycle in a given problem. Depending on times and demand growth rates in the planning horizon, the m_ns can be different for different cycles. Hence, any CEC policy can be characterized by $(N, \vec{M})$, where $\vec{M} = \{m_n; n = 1, 2, \ldots, N\}$. Figure 3.2 provides an illustration of using sequential-replenishment policy with $(N, \vec{M}) = (2, (2, 1))$.

Further, throughout this chapter, we require the following assumption to hold.

Assumption 3.1. $\dfrac{a_{11}}{a_{21}} \geq \dfrac{D_1(\tau_{k,n}, \tau_{k+1,n})}{D_2(\tau_{k,n}, \tau_{k+1,n})} \geq \dfrac{a_{12}}{a_{22}}$ *for any* $k \in \{1, 2, ..., m_n\}$ *of any cycle* $n \in \{1, ..., N\}$.

In Assumption 3.1, an *attribute-i-intense cluster* is referred to as cluster-type i ($i = 1, 2$). This assumption has a practical implication, as it assures that, with a combined use of different cluster-types, the cloud provider should be able to achieve a balance between supply and demand in each cycle. Otherwise, at replenishment times when Assumption 3.1 is not satisfied, employment of those clusters will result in the accumulation of excess capacities of one of the attributes, which is against the provider's desire to reduce excess capacities through the employment of those clusters. As noted, because the cloud provider can collaborate with its vendors on customizing the cluster-types at the beginning of the horizon, it is reasonable to assume that the configured cluster-types must satisfy Assumption 3.1.

Although Assumption 3.1 does not necessarily hold for arbitrary replenishment times, we will consider only policies for which the replenishment times are feasible and, thus, satisfy

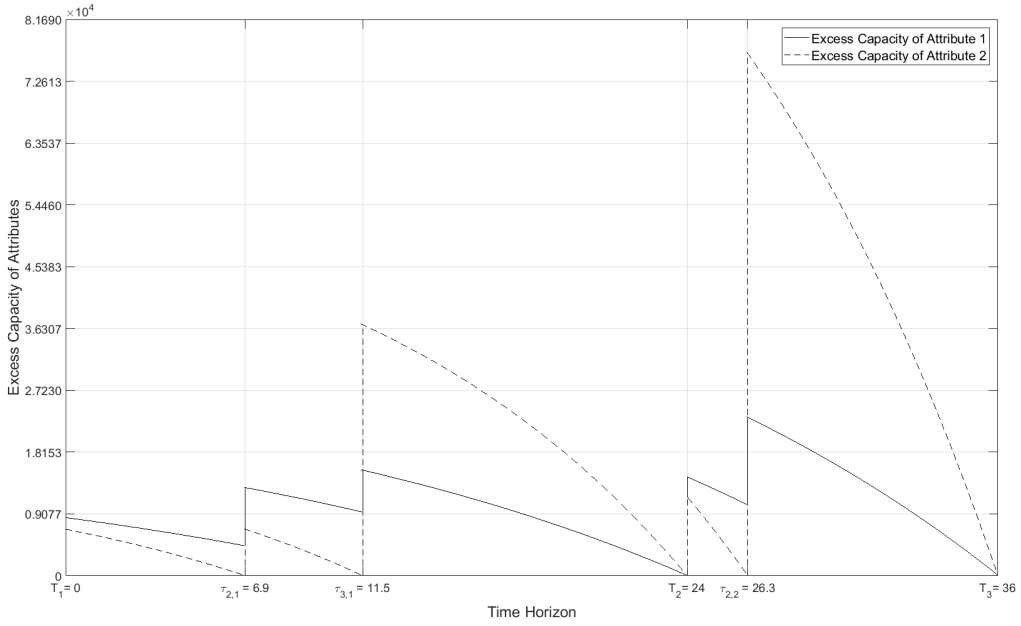


Figure 3.2: An illustration of the excess capacities of attributes 1 and 2 under a specific CEC policy.The setting of this example: $T = 36$ (months), demand growths are exponential ($D_i(t) = \gamma e^{\lambda_i t}$, where $\gamma = 10^5$, $\lambda_1 = 0.05$, and $\lambda_2 = 0.075$), Cluster 1 ($a_{11} = 50$ and $a_{21} = 40$) is the leading cluster and Cluster 2 ($a_{12} = 10$ and $a_{22} = 60$) is the following cluster. The specific policy consists of two capacity expansion cycles. In the first cycle, $[T_1 = 0, T_2 = 24]$, we purchase cluster 1 twice (with a batch size of 170) followed by a single replenishment of cluster 2 (with a batch size 615); in the second cycle, $[T_2 = 24, T_3 = 36]$, we purchase cluster 1 once (with a batch size of 290) followed by a single replenishment of cluster 2 (with a batch size 1278).

Assumption 3.1. For example, if the replenishment times correspond to fixed decision points, such as the beginning of every month, which is the regular pace that many firms follow for key capacity-expansion decisions, then the provider needs to ensure that the configured cluster-types satisfy Assumption 3.1 for those given replenishment times. If the provider also has the flexibility of determining the replenishment times, such as the possibility of changing the current pace, the provider needs to ensure that the configured cluster-types and the desired replenishment times together satisfy Assumption 3.1.

Next, we derive the operating characteristics of the general $(N, \vec{M})$ policies based on sequential- and simultaneous-replenishments in Sections 3.1.1 and 3.1.2, respectively.

3.1.1 Operating Characteristics of the Sequential-Replenishment Policies

For ease of exposition, we will first derive the total cost within any given CEC with cluster 1 ($j = 1$) being the leading cluster. The derivation of the total cost with cluster 2 being the leading cluster can be obtained in a similar vein. With a slight abuse of notation, for a cycle $n \in \{1, 2, \ldots, N\}$, let m be the number of replenishments of the leading cluster (i.e., $m = m_n$).

When the leading cluster is employed in the first m replenishments, the excess capacity of attribute 2 reduces faster than that of attribute 1 because a_{11}/a_{21} is greater than $D_1(\tau_{k,n}, \tau_{k+1,n})/D_2(\tau_{k,n}, \tau_{k+1,n})$ for any $k \in \{1, 2, ..., m\}$, according to Assumption 3.1. As a result, the excess capacity of attribute 1 is positive when attribute 2's excess capacity reaches zero, upon which the next replenishment has to be made. The excess capacity of attribute 1 accumulates over the cycle until the last replenishment. When cluster 2 is purchased at the last expansion of the cycle, there must be excess capacity of attribute 1 to start with, but then the excess capacity of attribute 1 will deplete faster than that of attribute 2 because a_{12}/a_{22} is smaller than $D_1(\tau_{m+1,n}, T_{n+1})/D_2(\tau_{m+1,n}, T_{n+1})$. Eventually, the capacities of both

attributes can be leveled at the same time at the end of this cycle. Thus, for attribute 2, we have

$$a_{21}Q_{k,1,n} = D_2(\tau_{k,n}, \tau_{k+1,n}), \text{ for } k = 1, \ldots, m, \text{ and } a_{22}Q_{m+1,2,n} = D_2(\tau_{m+1,n}, T_{n+1}). \quad (3.1)$$

For attribute 1, because its capacity should be leveled at the end of the cycle, we should have

$$a_{11}\sum_{k=1}^{m} Q_{k,1,n} + a_{12}Q_{m+1,2,n} = D_1(T_n, T_{n+1}). \quad (3.2)$$

Define $\alpha = a_{11}a_{22} - a_{12}a_{21}$. We obtain

$$\sum_{k=1}^{m} Q_{k,1,n} = \frac{a_{22}D_1(T_n, T_{n+1}) - a_{12}D_2(T_n, T_{n+1})}{\alpha}, Q_{m+1,2,n} = \frac{a_{11}D_2(T_n, T_{n+1}) - a_{21}D_1(T_n, T_{n+1})}{\alpha}.$$

Then, the total replenishment (purchase) cost will be

$$\begin{aligned} & c_1\sum_{k=1}^{m} Q_{k,1,n} + c_2Q_{m+1,2,n} \\ &= c_1\frac{a_{22}D_1(T_n, T_{n+1}) - a_{12}D_2(T_n, T_{n+1})}{\alpha} + c_2\frac{a_{11}D_2(T_n, T_{n+1}) - a_{21}D_1(T_n, T_{n+1})}{\alpha}, \end{aligned}$$

which depends on the start and the end times of the cycle but is independent of m.

Next, the holding cost of attribute 2 will be

$$h_2\left\{\sum_{k=1}^{m}\int_{\tau_{k,n}}^{\tau_{k+1,n}}[a_{21}Q_{k,1,n} - D_2(\tau_{k,n}, t)]dt + \int_{\tau_{m+1,n}}^{T_{n+1}}[a_{22}Q_{m+1,2,n} - D_2(\tau_{m+1,n}, t)]dt\right\}.$$

To derive the holding cost for attribute 1, note that at the k-th ($k = 1, \ldots, m$) replenishment in the cycle, the excess capacity of attribute 1 will be $a_{11}\sum_{l=1}^{k} Q_{l,1,n} - D_1(T_n, \tau_{k,n})$. Thus, the holding cost of attribute 1 from $\tau_{k,n}$ until $\tau_{k+1,n}$ (before the $(k+1)$-th replenishment

is placed) is

$$h_1 \left\{ \int_{\tau_{k,n}}^{\tau_{k+1,n}} \left[a_{11} \sum_{l=1}^{k} Q_{l,1,n} - D_1(T_n, \tau_{k,n}) - D_1(\tau_{k,n}, t) \right] dt \right\}.$$

Further, at the time when the last replenishment (i.e., the $(m+1)$-th replenishment) is made, the excess capacity of attribute 1 will be $a_{11} \sum_{k=1}^{m} Q_{k,1,n} + a_{12} Q_{m+1,2,n} - D_1(T_n, \tau_{m+1,n})$. Thus, employing (3.1), the holding cost of attribute 1 from $\tau_{m+1,n}$ until the end of this cycle will be

$$\begin{aligned}
&h_1 \left\{ \int_{\tau_{m+1,n}}^{T_{n+1}} \left[a_{11} \sum_{k=1}^{m} Q_{k,1,n} + a_{12} Q_{m+1,2,n} - D_1(T_n, \tau_{m+1,n}) - D_1(\tau_{m+1,n}, t) \right] dt \right\} \\
&= h_1 \left\{ \int_{\tau_{m+1,n}}^{T_{n+1}} \left[\frac{a_{11}}{a_{21}} \sum_{k=1}^{m} D_2(\tau_{k,n}, \tau_{k+1,n}) + a_{12} Q_{m+1,2,n} - D_1(T_n, \tau_{m+1,n}) - D_1(\tau_{m+1,n}, t) \right] dt \right\} \\
&= h_1 \left\{ \int_{\tau_{m+1,n}}^{T_{n+1}} \left[\frac{a_{11}}{a_{21}} D_2(T_n, \tau_{m+1,n}) + a_{12} Q_{m+1,2,n} - D_1(T_n, \tau_{m+1,n}) - D_1(\tau_{m+1,n}, t) \right] dt \right\}.
\end{aligned}$$

Lemma 3.1 below summarizes the cost terms in a given cycle. Define

$$F_i(a,b) = \int_a^b D_i(a,t)dt, i \in \{1,2\}, \text{ and } w_j = h_1 a_{1j} + h_2 a_{2j}, j \in \{1,2\}.$$

Lemma 3.1. *In a given cycle $[T_n, T_{n+1}]$ with cluster $l \in \{1,2\}$ as the leading cluster and cluster $f \neq l$ as the following cluster, the cost terms will be*

- *Total fixed expansion cost: $A(m+1)$.*

- *Total replenishment (purchase) cost is*

$$\frac{(c_1 a_{22} - c_2 a_{21})\, D_1(T_n, T_{n+1}) + (c_2 a_{11} - c_1 a_{12}) D_2(T_n, T_{n+1})}{\alpha}. \tag{3.3}$$

- *Total holding cost:*

$$\sum_{k=1}^{m}(\tau_{k+1,n}-\tau_{k,n})\frac{w_l D_f(T_n,\tau_{k+1,n})}{a_{fl}}+\psi_n, \tag{3.4}$$

where $\psi_n=\sum_{i=1}^{2}h_i[(T_{n+1}-\tau_{m+1,n})D_i(T_n,T_{n+1})-F_i(T_n,T_{n+1})]$ *is independent of* m.

Proof: See Appendix.

Lemma 3.1 implies that the total fixed expansion cost and the total replenishment cost are not affected by which cluster is the leading cluster. Thus, one needs to compare just the total holding costs with different leading clusters to determine which cluster-type should be the leading cluster.

3.1.2 Operating Characteristics of the Simultaneous-Replenishment Policies

In contrast to the sequential-replenishment policy, the cloud provider can employ both cluster-types jointly in each replenishment. As in the previous section, we consider $m+1$ replenishment opportunities in the n-th CEC with the replenishment times $\tau_{k,n}$, where $k\in\{1,2,...,m+1\}$. Moreover, the cloud provider can achieve a balance between supply and demand in every joint replenishment, which is feasible due to Assumption 3.1. That is,

$$a_{11}Q_{k,1,n}+a_{12}Q_{k,2,n}=D_1(\tau_{k,n},\tau_{k+1,n}),\ k=1,2,...,m+1.$$
$$a_{21}Q_{k,1,n}+a_{22}Q_{k,2,n}=D_2(\tau_{k,n},\tau_{k+1,n}),\ k=1,2,...,m+1.$$

Thus,

$$Q_{k,1,n}=\frac{a_{22}D_1(\tau_{k,n},\tau_{k+1,n})-a_{12}D_2(\tau_{k,n},\tau_{k+1,n})}{\alpha},\ k=1,2,...,m+1.$$
$$Q_{k,2,n}=\frac{a_{11}D_2(\tau_{k,n},\tau_{k+1,n})-a_{21}D_1(\tau_{k,n},\tau_{k+1,n})}{\alpha},\ k=1,2,...,m+1.$$

Then, the total replenishment (purchase) cost will be

$$c_1 \sum_{k=1}^{m+1} Q_{k,1,n} + c_2 \sum_{k=1}^{m+1} Q_{k,2,n}.$$

Next, the total holding cost of attribute $i \in \{1, 2\}$ will be

$$h_i \left\{ \sum_{k=1}^{m+1} \int_{\tau_{k,n}}^{\tau_{k+1,n}} [a_{i1} Q_{k,1,n} + a_{i2} Q_{k,2,n} - D_i(\tau_{k,n}, t)] dt \right\}. \tag{3.5}$$

Lemma 3.2 below summarizes the cost terms in the given cycle.

Lemma 3.2. *In a given cycle $[T_n, T_{n+1}]$ with joint replenishments of both cluster-types in each expansion, the cost terms will be*

- *Total fixed expansion cost: $\beta A(m+1)$, where $\beta \in [1, 2]$.*

- *Total replenishment (purchase) cost is:*

$$\frac{(c_1 a_{22} - c_2 a_{21})\, D_1(T_n, T_{n+1}) + (c_2 a_{11} - c_1 a_{12}) D_2(T_n, T_{n+1})}{\alpha}. \tag{3.6}$$

- *Total holding cost:*

$$\sum_{i=1}^{2} \left\{ h_i \left[\sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) D_i(T_n, \tau_{k+1,n}) - F_i(T_n, T_{n+1}) \right] \right\}. \tag{3.7}$$

Proof: See Appendix.

Before proceeding further, a discussion of a few implications of Lemma 3.2 is in order.

First, the parameter $\beta \in [1, 2]$ captures the potential savings on the fixed expansion costs due to the joint replenishment of both cluster-types in each expansion. In particular, $\beta = 1$ implies that the fixed expansion cost of employing both cluster-types in each replenishment

is equal to that of employing a single cluster-type, while $\beta = 2$ implies that the fixed expansion cost of employing both cluster-types in each replenishment is twice as much as that of employing a single cluster-type. As β increases, the savings on the fixed expansion costs decrease.

Second, the total replenishment (purchase) cost under the simultaneous-replenishment policy is the same as that under the sequential-replenishment policy, as the added capacity in any cycle must be equal to the growth of demand for each attribute.

Third, for any given cycle with fixed expansion times, we can easily compare the cost terms between the sequential- and the simultaneous-replenishment policies, using Lemmas 3.1 and 3.2. It is obvious that, *with the expansion times fixed*, the total fixed expansion cost under the simultaneous-replenishment policy is greater than or equal to that under the sequential-replenishment policy, as the former (latter) requires replenishment of both (only one) cluster-types (cluster-type) at each expansion. Moreover, we can show that, *with the expansion times fixed*, the total holding cost under the simultaneous-replenishment policy is smaller than that under the sequential-replenishment policy. The proof of this result is part of the proof for Proposition 3.1.

Fourth, in contrast to the holding cost under sequential-replenishment policy, the holding cost under simultaneous-replenishment policy is independent of cluster configurations (i.e., a_{ij}s). This implies, if the cloud firm intends to find the optimal pair of clusters under simultaneous-replenishment policy, it should find the pair with minimum replenishment cost.

Note that the third remark on the comparison between the sequential- and simultaneous-replenishment policies is based on a condition that the expansion times are fixed. Nevertheless, the optimal sequential- and simultaneous-replenishment policies may have different expansion times, so the comparison of the total costs between these optimal policies is not trivial. Section 3.1.3 below provides the final result of such a comparison.

3.1.3 Comparison Between the Sequential- and Simultaneous-Replenishment Policies

Proposition 3.1. *There exists a unique threshold* $\overline{\beta} \in (1, 2)$ *such that the total cost of the optimal simultaneous-replenishment policy is smaller than that of the optimal sequential-replenishment policy if and only if* $\beta \leq \overline{\beta}$.

Proof: See Appendix.

The proof for Proposition 3.1 is based on the results below. First, when $\beta = 1$, the total cost of the optimal simultaneous-replenishment policy is strictly lower than that of the optimal sequential-replenishment policy, as for any given sequential-replenishment policy, we can construct a simultaneous-replenishment policy with the same expansion times. In such a way, the total fixed expansion costs of these policies are the same, while the total holding cost of the simultaneous-replenishment policy is strictly lower than that of the given sequential-replenishment policy.

Second, when $\beta = 2$, the total cost of the optimal simultaneous-replenishment policy is strictly higher than that of the optimal sequential-replenishment policy, as for any given simultaneous-replenishment policy, we can find a sequential-replenishment policy, which has a single expansion of a cluster-type followed by a single expansion of the other cluster-type between any two consecutive expansion times of the given simultaneous-replenishment policy. In such a way, the total fixed expansion costs of these policies are the same, while the total holding cost of the sequential-replenishment policy is strictly lower than that of the given simultaneous-replenishment policy.

Finally, it is obvious that the total cost of the optimal simultaneous-replenishment policy is an increasing function of $\beta \in [0, 1]$. Therefore, there must exist a unique threshold, $\overline{\beta} \in (1, 2)$, such that the total costs of the optimal simultaneous- and sequential-replenishment policies are equal.

In summary, Proposition 3.1 implies that, if the savings on the fixed expansion costs of the simultaneous-replenishment policy are significant (or negligible), the cloud provider should consider the employment of different cluster-types simultaneously (sequentially).

3.2 Analysis of the Replenishment Policies for One Cycle

Having derived the total cost of the sequential- and simultaneous-replenishment policies in Section 3.1, we now study the optimization of those policies over one CEC. To obtain deeper analysis, we study cases with realistic demand growth patterns and under practical replenishment policies.

First, we assume that the growth in demand for each attribute is exponential with time. Such a growth pattern is suitable in today's environment, where the demand for cloud applications is rapidly growing over time. For example, Figure 3.3 shows cloud traffic growth, in different regions of the world, is exponentially increasing[3]. In fact, according to a survey by Cisco (Cisco 2016), the projected cloud demand follows an exponential growth rate with a compound annual growth rate of around 24.0%. The exponential demand growth also has been commonly studied in the operations management literature on capacity expansions (Luss 1982), and our study is in line with that literature. Specifically, let $D_i(a,b) = \gamma(e^{\lambda_i b} - e^{\lambda_i a})$ (for any times $a < b$ within $[0,T]$) be the growth of demand for attribute i from times a to b, where $\gamma > 0$ controls the scale and $\lambda_i > 0$ ($i \in \{1,2\}$) controls the growth rate of demand for attribute i.

Second, note that in practice often:

1. The intervals between consecutive replenishment times are equal. As an example for this case, the replenishment times may correspond to the beginning of every month, which is the pace of the cloud provider's capacity-expansion decisions.

[3]https://www.businessinsider.com/alibaba-enters-the-us-cloud-computing-fray-2015-6

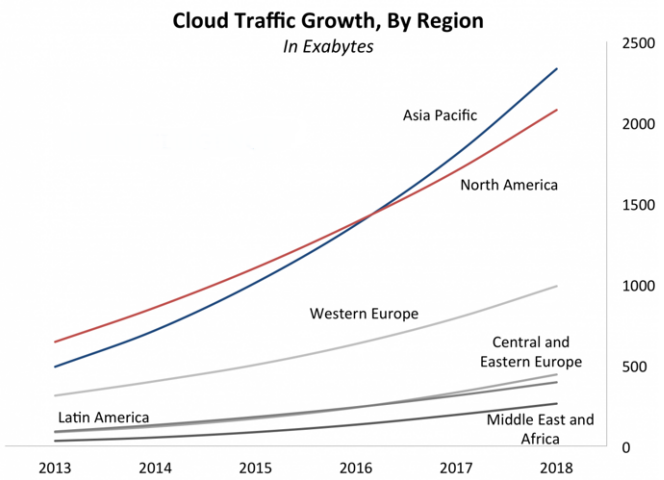


Figure 3.3: Cloud traffic growth in different regions of the world

2. Replenishments within a cycle have the same batch size. As an example for this case, in the contract, cloud company may require the supplier to deliver 50000 servers in every replenishment.

Hence, we study two replenishment policies: 1) equal-interval policy, and 2) equal-batch policy.

3.2.1 Analysis of the Sequential-Replenishment Policies

Without loss of generality, we analyze the case where cluster 1 is the leading cluster, as the other case can be analyzed in the same vein. That is, under the sequential-replenishment policy, the cloud provider employs cluster 1 at replenishment time $\tau_{k,n}$, where $k = 1, ..., m$, and cluster 2 at replenishment time $\tau_{m+1,n}$. Let $t_{k,n} = \tau_{k+1,n} - \tau_{k,n}$ for any $k \in \{1, ..., m+1\}$. Here, we keep the subscript n to facilitate an extension to the general model (Section 3.3.1), for which the planning horizon consists of multiple cycles. The reader should be mindful that, in Sections 3.2.1 and 3.2.2, $\tau_{1,n} = T_n = 0$ and $\tau_{m+2,n} = T_{n+1} = T$, as the planning horizon now consists of only one cycle.

Moreover, note that under the sequential-replenishment policy, $\tau_{m+1,n}$ is independent of m. Indeed, $\tau_{m+1,n}$ is uniquely determined by

$$\frac{a_{11}}{a_{21}} D_2(T_n, \tau_{m+1,n}) + \frac{a_{12}}{a_{22}} D_2(\tau_{m+1,n}, T_{n+1}) = D_1(T_n, T_{n+1}). \tag{3.8}$$

To derive (3.8), note that $a_{21}Q_{k,1,n} = D_2(\tau_{k,n}, \tau_{k+1,n})$ for $k = 1, \ldots, m$, $a_{22}Q_{m+1,2,n} = D_2(\tau_{m+1,n}, T_{n+1})$, and $\sum_{k=1}^{m} a_{11}Q_{k,1,n} + a_{12}Q_{m+1,2,n} = D_1(T_n, T_{n+1})$, according to (3.1) and (3.2). Because $\tau_{m+1,n}$ is fixed, the interval of the last expansion, $t_{m+1,n} = T_{n+1} - \tau_{m+1,n}$, is also fixed. In what follows, we study equal-interval and equal-batch policies.

1. If we impose the condition that all intervals must be equal over the entire cycle, we should have $t_{k,n} = t_{m+1,n}$ for all $k = 1, ..., m$. Such an equal-interval policy is feasible, however, only if the length of the given cycle is a multiple of $t_{m+1,n}$. Otherwise, we propose that the policy requires $t_{k,n}$, where $k = 1, ..., m$, to be constant, but the constant interval is not required to be equal to the last interval, $t_{m+1,n}$. We believe that this policy is also practical, as it is basically an equal-interval policy when the cloud provider deploys the leading clusters, while the interval of the last replenishment can be different when the provider needs to deploy the following cluster. Henceforth, we will refer to the sequential-replenishment policy, under the condition that all intervals must be equal, as the *Equal-Interval Policy with Sequential Replenishment* or the *EI-Sequential* policy. We will refer to the proposed policy as the *m-Equal-Interval Policy with Sequential Replenishment* or the *m-EI-Sequential* policy. It is worth noting that the cloud provider needs to find the optimal m under the m-EI-Sequential policy, while the provider does not need to make a decision under the EI-Sequential policy. Moreover, the EI-Sequential policy, when it is feasible, is a special case of the m-EI-Sequential policy. Thus, the optimal m-EI-Sequential policy must perform better than

the EI-Sequential policy.

2. Recall that under sequential-replenishment policy, at expansion times when there is a replenishment of cluster 1, there will be excess capacity of attribute 1, while the capacity of attribute 2 will be leveled. Then, in the last expansion of a CEC, cluster 2 is replenished at $\tau_{m+1,n}$ to balance the demand and supply of both attributes at T_{n+1}. Under equal-batch policy, we assume all first m expansions have the same expansion quantity. That is, $ka_{21}Q_{1,n} = D_2(T_n, \tau_{k+1,n})$ for $k = 1, 2, ..., m$. Therefore,

$$D_2(T_n, \tau_{k+1,n}) = \frac{k}{m} D_2(T_n, \tau_{m+1,n}). \tag{3.9}$$

We refer to this policy as *EB-Sequential* policy. Next, $\tau_{k+1,n}$ can be easily derived from (3.8) and (3.9) with respect to problem parameters.

We now focus on optimization of the m-EI-Sequential and EB-Sequential policies. The objective is to choose m to minimize the total cost excluding the replenishment cost as it is a constant over the planning horizon. According to Lemma 3.1, the total cost, under m-EI-Sequential policy is

$$A(m+1)+\sum_{k=1}^{m}(\tau_{k+1,n}-\tau_{k,n})\frac{w_1 D_2(T_n, \tau_{k+1,n})}{a_{21}}+\psi_n = A(m+1)+\frac{w_1 t_n}{a_{21}}\sum_{k=1}^{m} D_2(T_n, T_n+kt_n)+\psi_n, \tag{3.10}$$

where $t_n = (\tau_{m+1,n} - T_n)/m$ and $\tau_{m+1,n}$ is determined by (3.8). In contrast, under the EB-Sequential policy, the objective function reduces to

$$A(m+1) + \frac{w_1}{a_{21}}\left(\tau_{m+1,n} - \sum_{k=1}^{m}\frac{\tau_{k,n}}{m}\right) D_2(T_n, \tau_{m+1,n}) + \psi_n. \tag{3.11}$$

Further, when the growth in demand is exponential, we can obtain $\tau_{m+1,n}$ in closed form.

Lemma 3.3. *When $D_i(T_n, t) = \gamma(e^{\lambda_i t} - e^{\lambda_i T_n})$ for any $t \geq T_n$, the last expansion time of the cycle, $\tau_{m+1,n}$, is equal to $T_n + (1/\lambda_2)(\ln \chi_n)$, where χ_n is defined as*

$$\chi_n = e^{-\lambda_2 T_n} \left[\frac{a_{21}(a_{22}e^{\lambda_1 T_{n+1}} - a_{12}e^{\lambda_2 T_{n+1}}) + a_{22}(a_{11}e^{\lambda_2 T_n} - a_{21}e^{\lambda_1 T_n})}{\alpha} \right].$$

Proof: See Appendix.

As a result of Lemma 3.3, $t_{m+1,n} = T_{n+1} - \tau_{m+1,n} = T_{n+1} - T_n - (\ln \chi_n)/\lambda_2$. When the EI-Sequential policy is feasible, the constant interval between consecutive replenishments is equal to $t_{m+1,n}$, and thus $m = (T_{n+1} - T_n)/t_{m+1,n} = (T_{n+1} - T_n)/(T_{n+1} - T_n - (\ln \chi_n)/\lambda_2)$. By contrast, under the EB-Sequential and m-EI-Sequential policy, the provider needs to decide on m to minimize the total cost.

Proposition 3.2. *When $D_i(T_n, t) = \gamma(e^{\lambda_i t} - e^{\lambda_i T_n})$ for any $t \geq T_n$, the objective is to determine m to minimize the following function of the total cost:*

- ***If one employs the m-EI-Sequential policy,***

$$A(m+1) + \frac{w_1 \gamma e^{\lambda_2 T_n} \ln \chi_n}{a_{21}\lambda_2} \left(\frac{\sum_{k=1}^{m} \chi_n^{\frac{k}{m}}}{m} - 1 \right) + \psi_n, \tag{3.12}$$

- ***If one employs the EB-Sequential policy***

$$A(m+1) + \frac{\gamma w_1 e^{\lambda_2 T_n}}{a_{21}\lambda_2}(\chi_n - 1) \left(\ln \chi_n - \frac{\sum_{k=1}^{m-1} \ln\left(\frac{k}{m}\chi_n + (1 - \frac{k}{m})\right)}{m} \right) + \psi_n, \tag{3.13}$$

where χ_n (defined in Lemma 3.3) and ψ_n (defined in Lemma 3.1) are independent of m. Further, (3.12) is a convex function of m.

Proof: See Appendix.

Proposition 3.2 implies that 1) the optimal number of replenishments of the leading cluster under m-EI-sequential policy can be determined by applying standard convex optimization techniques. However, because the optimal number of replenishments of the leading cluster does not have a closed-form solution, one needs to estimate it from (3.12) numerically. 2) Since the total cost under EB-Sequential policy is not necessarily convex in m, the optimal number of replenishments of the leading cluster can be found numerically using (3.13).

Alternatively, one can approximate (3.10) and (3.11) by applying the first-order Taylor expansion of the exponential demand function. That is, $D_i(T_n, t) = \gamma e^{\lambda_i T_n}(e^{\lambda_i(t-T_n)} - 1) \approx \gamma e^{\lambda_i T_n}\lambda_i(t - T_n)$. Define $\lambda_{in} = \gamma\lambda_i e^{\lambda_i T_n}$; then, $D_i(T_n, t)$ can be approximated by $\lambda_{in}(t - T_n)$. Replacing $D_i(T_n, t) = \lambda_{in}(t - T_n)$ into (3.10) and solving that minimization problem, we can approximate the optimal number of replenishments of the leading cluster, m^*, in closed form. Before we proceed to the next proposition, note that using the approximated total demand expression, the difference between the m-EI-sequential policy and EB-Sequential policy no longer exists because they are equivalent under the constant incremental demand growth rates.

Proposition 3.3. *Using the approximation $D_i(T_n, t) \approx \gamma\lambda_i e^{\lambda_i T_n}(t - T_n) = \lambda_{in}(t - T_n)$, the approximate optimal number of replenishments of the leading cluster (e.g., cluster 1) is given by*

$$\widehat{m} = \sqrt{\frac{w_1 a_{21}}{2A\lambda_{2n}}}\frac{(\lambda_{1n}a_{22} - \lambda_{2n}a_{12})(T_{n+1} - T_n)}{\alpha}. \tag{3.14}$$

Proof: See Appendix.

We make two remarks on Proposition 3.3. First, in this proposition, $\widehat{m}$ is expressed as a real number, but one should enforce the integrality of $\widehat{m}$ by examining the two nearby integers of $\widehat{m}$. Second, while obtaining m^* based on the exact functional form of the exponential demand growth is more accurate, the alternative approximate approach results in a closed-form solution and thus has its own merits and advantages: (i) the approximate solution is

easy to communicate and convenient for the purpose of planning; and (ii) when the length of the cycle is not too long and the rates of the exponential demand functions are not too large, the approximate solution results in a total cost that is close enough to the optimum, and thus $\widehat{m}$ can be used as a good starting point for the search of m^*. For instance, through a numerical study whose setting is later described in Section 3.4.1, the increase in the total cost of using the approximate solution compared to that of using the exact optimal solution is reported in Table 3.1.

Table 3.1: Relative increased cost of using the approximated solution compared to exact optimal solution.

Minimum	1st Quartile	2nd Quartile	3rd Quartile	Maximum
0.00%	0.00%	0.27%	0.75%	1.15%

3.2.2 Analysis of the Simultaneous-Replenishment Policies

Under simultaneous-replenishment policies, it is always feasible to impose the constraint that 1) all replenishment intervals must be equal, or 2) all replenishment quantities must be equal. The reason is that the provider can achieve a supply-demand balance at each expansion through the simultaneous replenishment of both cluster-types. Next, we focus on equal-interval and equal-batch policies.

1. Under equal-interval policy, let $t_n = (T_{n+1} - T_n)/(m+1)$. The provider should employ both cluster types at replenishment times, $\tau_{k,n}$, where $\tau_{k+1,n} - \tau_{k,n} = t_n$ for $k = 1, ..., m+1$. Henceforth, we will refer to such a policy as the *Equal-Interval Policy with Simultaneous Replenishment* or the *EI-Simultaneous* policy. According to Lemma 3.2, the total cost of the EI-Simultaneous policy is as follows (excluding the total

replenishment/purchase cost as it is a constant over the planning horizon):

$$\begin{aligned} & \beta A(m+1) + \sum_{i=1}^{2} \left\{ h_i \left[\sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) D_i(T_n, \tau_{k+1,n}) - F_i(T_n, T_{n+1}) \right] \right\} \\ = \; & \beta A(m+1) + t_n \sum_{i=1}^{2} \left\{ h_i \sum_{k=1}^{m+1} D_i(T_n, T_n + kt_n) \right\} - \sum_{i=1}^{2} h_i F_i(T_n, T_{n+1}). \end{aligned} \tag{3.15}$$

2. Under equal-batch policy, $D_i(\tau_{k,n}, \tau_{k+1,n}) = \frac{1}{m+1} D_i(T_n, T_{n+1})$. Therefore, $D_i(T_n, \tau_{k+1,n}) = \frac{k}{m+1} D_i(T_n, T_{n+1})$. We refer to this policy as EB-Simultaneous policy. According to Lemma 3.2, the total cost of the EB-Simultaneous policy is as follows (excluding the total replenishment/purchase cost as it is a constant over the planning horizon):

$$\gamma A(m+1) + \sum_{i=1}^{2} \frac{h_i D_i(T_n, T_{n+1})}{m} \left[\sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) k + F_i(T_n, T_{n+1}) \right] \tag{3.16}$$

Proposition 3.4. *When $D_i(T_n, t) = \gamma(e^{\lambda_i t} - e^{\lambda_i T_n})$ for any $t \geq T_n$, the objective is to determine m to minimize the following function of the total cost:*

- ***If one employs EI-Sequential policy,***

$$\beta A(m+1) + \gamma\nu \sum_{i=1}^{2} \left\{ h_i e^{\lambda_i T_n} \left(\frac{\sum_{k=1}^{m+1} (e^{\lambda_i \nu})^{\frac{k}{m+1}}}{m+1} - 1 \right) \right\} + \phi_n, \tag{3.17}$$

- ***If one employs EB-Sequential policy,***

$$\beta A(m+1) + \sum_{i=1}^{2} \frac{h_i e^{\lambda_i T_n} (e^{\nu\lambda_i} - 1)}{m} \left[\sum_{k=0}^{m} \left(\nu - \frac{1}{\lambda_i} ln \left(\frac{k}{m} e^{\nu\lambda_i} + (1 - \frac{k}{m}) \right) \right) \right] + \phi_n \tag{3.18}$$

where $\nu = T_{n+1} - T_n$ and $\phi_n = -\sum_{i=1}^{2} h_i F_i(T_n, T_{n+1})$ is independent of m. Moreover, (3.17) is a convex function of m.

Proof: See Appendix.

Proposition 3.4 implies that 1) the optimal number of replenishments under EI-Simultaneous policy can be determined by applying standard convex optimization techniques. However, because the optimal number of replenishments does not have a closed-form solution, one needs to estimate it from (3.17) numerically. 2) Since the total cost under EB-Simultaneous policy is not necessarily convex in m, the optimal number of replenishments can be searched numerically using (3.18).

Alternatively, similar to the approximate approach of Section 3.2.1, we can replace $D_i(T_n, t) = \lambda_{in}(t - T_n)$, where $\lambda_{in} = \gamma\lambda_i e^{\lambda_i T_n}$, into (3.15). Then, we can derive the approximate optimal number of replenishments, $\widehat{m}$, in closed form.

Proposition 3.5. *Using the approximation* $D_i(T_n, t) \approx \gamma\lambda_i e^{\lambda_i T_n}(t - T_n) = \lambda_{in}(t - T_n)$, *the approximate optimal number of joint replenishments is given by*

$$\widehat{m} = (T_{n+1} - T_n)\sqrt{\frac{\sum_{i=1}^{2} h_i\lambda_{in}}{2\beta A}} - 1. \tag{3.19}$$

Proof: See Appendix.

The remarks on Proposition 3.3 also apply to Proposition 3.5. Furthermore, Table 3.2 reports the statistic summary of the relative increased cost of using the approximated solution compared to that of using the exact optimal solution.

Table 3.2: Relative increased cost of using the approximated solution compared to exact optimal solution.

	Minimum	1st Quartile	2nd Quartile	3rd Quartile	Maximum
$\beta = 1$	0.00%	0.00%	0.30%	0.49%	1.21%
$\beta = 1.5$	0.00%	0.00%	0.30%	0.48%	1.62%
$\beta = 2$	0.00%	0.00%	0.30%	0.77%	0.95%

3.3 Extensions

In this section, we present three extensions of our base model. First, we consider the possibility that the provider may want to divide the planning horizon into *multiple CECs*. Second, we investigate the optimal configurations of the two cluster-types at the beginning of each CEC, which we will refer to as the provider's *cluster-selection problem*. Third, we study the case where multiple demand scenarios are forecasted by the forecasting team.

3.3.1 Analysis of the Replenishment Policies for Multiple Cycles

We now consider the possibility that the provider wants to divide the planning horizon into $N \geq 1$ cycles. As an example, the planning horizon can be one year, but the provider aims to achieve a balance between supply and demand (i.e., the capacities of both attributes are leveled) every six months. In this example, $N = 2$ and each cycle has a span of six months.

It is also worth noting that, due to the rapid development of computer science technologies, the provider may be able to re-customize the cluster-types at certain times. The re-customization, however, cannot occur too often, as designing and testing a new cluster-type takes a significant amount of time (six months to one year). Hence, dividing the planning horizon into multiple CECs enables the provider to change the selection of the cluster-types at the beginning of each cycle, while the selected cluster-types should be employed throughout the entire cycle. In the example above, where T is one year and $N = 2$, the provider can change the selection of the cluster-types every six months, while still employing the selected cluster-types within each cycle. We will study the provider's cluster-selection problem later (Section 3.3.2).

The objective is to determine the lengths of the CECs to minimize the total cost of the entire planning horizon using approximated demand growth. We first provide a dynamic-programming-based algorithm, namely, the *DP-algorithm*, to find the optimal lengths of the

CECs. Then, we develop a forward-looking heuristic, namely, the *FL-heuristic*, to achieve the same objective.

The DP-algorithm: Let $TC(T_n, T_{n+1})$ be the optimal total cost of a cycle starting at T_n and ending at T_{n+1} *for the approximated demand growth*. If we use the exact exponential demand growth, Proposition 3.2 provides the objective function of this one-cycle problem; given this, we need to numerically evaluate the m^* and the optimal cost. By contrast, if we use the approximate approach, Proposition 3.3 provides the optimal total cost in closed-form. Because the replenishment cost of the entire planning horizon is constant and independent of the decision parameters, we can exclude it from the calculation of the optimal total cost below.

Let $V_n(T_n)$ $(n = 1, 2, \ldots, N)$ denote the optimal total cost from T_n until the end of the planning horizon, $T_{N+1} = T$. When $N = 1$, the planning horizon contains only one cycle, and the optimization problem reduces to that of Section 3.2. When $N \geq 2$, the DP can be implemented using backward induction. Note that $V_N(T_N) = TC(T_N, T)$, and, recursively, for $n = N - 1, \ldots, 1$

$$V_n(T_n) = \min_{T_{n+1} \in (T_n, T]} TC(T_n, T_{n+1}) + V_{n+1}(T_{n+1}). \tag{3.20}$$

Finally, $V_1(T_1 = 0)$ gives the optimal total cost when there are N CECs. Then, we need to repeat the above procedure for $N = 3, 4, 5$, *etc.*, to obtain the optimal number of CECs, N^*. Although it is difficult to prove that the total cost of the entire planning horizon is a unimodal function of N, we have observed such a property in our numerical study.

The DP-algorithm determines the exact optimal number of CECs as well as their lengths. The computation, however, can be tedious for large values of N^*. Another disadvantage of the DP-algorithm is that the solution is not transparent and is difficult to explain to the managers. Thus, in addition to developing the exact DP-algorithm, we also propose the

following heuristic.

The FL-heuristic: We devise a forward-looking method, which myopically determines the end time of each CEC by minimizing the total cost rate of that cycle only. That is, at T_n, we determine T_{n+1} such that the total cost rate of this cycle, $TC(T_n, T_{n+1})/(T_{n+1} - T_n)$, is minimized.

To implement the heuristic, the first step is at $T_1 = 0$, find T_2 that minimizes $TC(T_1, T_2)/(T_2 - T_1)$. Then, move on to T_2, and find T_3 that minimizes $TC(T_2, T_3)/(T_3 - T_2)$. Repeat the procedure until at some N, $T_{N+1} > T$. Let $T_{N+1} = T$, and the heuristic ends.

As noted in Section 3.2, we can use the approximate approach to deal with the exponential demand growth. If one adopts the approximate approach to estimate $TC(T_n, T_{n+1})$, the proposition below assures that the optimal end time of the cycle starting from T_n, namely, $\widehat{T}_{n+1}$, which minimizes the total cost rate of the cycle, exists and can be easily found.

Proposition 3.6. *Using the approximation* $D_i(T_n, t) \approx \gamma\lambda_i e^{\lambda_i T_n}(t - T_n) = \lambda_{in}(t - T_n)$, *the total cost rate of a cycle starting from* T_n, $TC(T_n, T_{n+1})/(T_{n+1} - T_n)$, *is a convex function of* T_{n+1}.

Proof: See Appendix.

We believe that the proposed FL-heuristic is valuable for the following reasons: (i) it is easy to communicate due to the intuitive interpretation of the procedure (i.e., to minimize the total cost rate of each cycle), which is a distinct advantage over the non-transparent DP-algorithm; (ii) the forward-looking nature of the heuristic can be extremely important in real-world implementation as the heuristic solution can be easily modified to deal with the rolling horizon effect; and (iii) in Section 3.4.2, we will demonstrate that the heuristic performs very well, and, hence, its solution can be used as an excellent starting point for the DP-algorithm to reduce computational time.

3.3.2 The Cluster-Selection Problem

In the previous sections, we focused on the provider's optimal capacity-expansion plan, given two types of pre-configured cluster-types. The next question is: How should the provider select the two cluster-types from a collection of possible configurations? In this section, we first examine the provider's cluster-selection problem with a focus on two cluster-types using equal-interval policy as it is more common in practice compared to equal-batch policy (moreover, our results in this section can be easily extended to equal-batch policy). Then, we will consider the possibility of selecting multiple cluster-types and answer the following question: When is the selection of two cluster-types as good as the selection of multiple cluster-types?

3.3.2.1 The cluster-selection problem under the sequential-replenishment policies.

At the beginning of each CEC, the provider can conduct an exhaustive search for the optimal pair of cluster-types from a variety of possible configurations. For each pair that is being used, we can estimate the total cost of the cycle, applying the results of Section 3.2.1. Then, the optimal pair can be picked by comparing the total costs of all possible pairs. Although the exhaustive search results in the exact optimal selection, it can be time consuming when there are numerous possible configurations, and the exhaustive search does not provide insights into the optimal selection.

To address the disadvantages of the exhaustive search, we propose an efficient and intuitive cluster-selection heuristic based on the following: (i) select a cluster-type as the leading cluster, whose configuration would result in the minimum excess capacity at expansions with the leading cluster's being employed; and (ii) select another cluster-type as the following cluster, whose configuration would result in the fastest depletion of the excess capacity accumulated in the previous expansions. This heuristic is formally stated by Definition 3.1

below. Recall that $\lambda_{in} = \gamma e^{\lambda_i T_n}$ $(i \in \{1, 2\})$ is the approximate demand growth rate at the beginning of any cycle n.

Definition 3.1. ***(The cluster-selection heuristic).*** *Let $l, f \in \{1, 2\}$ and $l \neq f$. First, select the leading cluster by minimizing a_{ll}/a_{fl} among all possible attribute-l-intense clusters, where $a_{ll}/a_{fl} > \lambda_{ln}/\lambda_{fn}$. Then, select the following cluster by minimizing a_{lf}/a_{ff} among all possible attribute-f-intense clusters, where $a_{lf}/a_{ff} < \lambda_{ln}/\lambda_{fn}$.*

Consider the following *illustrative example*, where an attribute-1-intense cluster will be employed as the leading cluster. Suppose that there are six possible cluster-types, ranked as follows:

$$\frac{a''_{11}}{a''_{21}} > \frac{a'_{11}}{a'_{21}} > \frac{a_{11}}{a_{21}} > \frac{\lambda_{1n}}{\lambda_{2n}} > \frac{a_{12}}{a_{22}} > \frac{a'_{12}}{a'_{22}} > \frac{a''_{12}}{a''_{22}}.$$

The cluster-selection heuristic suggests selecting the cluster with parameters a_{11} and a_{21} as the leading cluster, as well as the cluster with parameters a''_{12} and a''_{22} as the following cluster. Intuitively, employing the selected leading cluster would result in the lowest excess capacity of attribute 1 among all the attribute-1-intense clusters. Then, employing the selected following cluster would deplete the excess capacity of attribute 1 the fastest among all the attribute-2-intense clusters.

Whether an attribute-1- or attribute-2-intense cluster, however, will be employed as the leading cluster can alter the suggestion made by the cluster-selection heuristic. We refer to such a phenomenon as the *flip-flop of the cluster-selection heuristic solution*. Consider again the above-mentioned illustrative example, except that an attribute-2-intense cluster is now employed as the leading cluster. Note that those available cluster-types can be rearranged

as follows:

$$\frac{a''_{22}}{a''_{12}} > \frac{a'_{22}}{a'_{12}} > \frac{a_{22}}{a_{12}} > \frac{\lambda_{2n}}{\lambda_{1n}} > \frac{a_{21}}{a_{11}} > \frac{a'_{21}}{a'_{11}} > \frac{a''_{21}}{a''_{11}}.$$

The cluster-selection heuristic now suggests choosing the cluster with parameters a_{12} and a_{22} as the leading cluster, as well as the cluster with parameters a''_{11} and a''_{21} as the following cluster. Due to the flip-flop of the cluster-selection heuristic solution, we should compare the total costs of the two different alternatives suggested by the heuristic and choose the one with the lower total cost.

The heuristic does not always lead to the exact optimal selection of cluster-types, which can be determined by the exhaustive search, as the heuristic relies on the approximate demand growth rates, λ_{in}, in classifying the possible configurations into two categories. As we noted earlier, however, the error of the approximation is insignificant if the length of the cycle is not very long or the λ_{in}'s are not very large. In reality, a cycle is usually not very long, as the planning horizon typically spans one year, and the planning horizon itself can be divided into multiple cycles.

We make a final remark about the cluster-selection heuristic. It is important to note that the cluster-selection heuristic has provided two desirable **ratios**, a_{1i}/a_{2i} where $i = 1, 2$, for the two cluster-types to be selected, but there could be multiple cluster-types that all satisfy the same ratio. In such a case, one needs to further consider the impact of the selected configuration on the total replenishment (purchase) cost. Specifically, if the unit purchase cost of a cluster-type is a weighted sum of its attributes, then the total replenishment cost is independent of the selected configuration. In general, however, the unit purchase cost of a cluster-type can be a concavely or convexly increasing function of the amounts of CPU and RAM bundled in one unit of the cluster-type, implying the economies or diseconomies of

scale, respectively. Then, among all cluster-types that satisfy the desired ratio of attributes, one should choose the cluster-type with the greatest (smallest) a_{ij}s, if the unit purchase cost exhibits the economies (diseconomies) of scale.

3.3.2.2 The cluster-selection problem under the simultaneous-replenishment policies.

When the provider implements a simultaneous-replenishment policy, the cluster-selection problem becomes relatively straightforward. The reason is, regardless of the selected cluster-types, one can always bring the capacities of the two attributes to the desired levels through the simultaneous replenishment of the selected cluster-types right after each expansion. As a result, the holding and the fixed expansion costs will be independent of the provider's selection of cluster-types, and the problem boils down to a comparison of the total replenishment costs of different selections. Note that this line of reasoning relies on only one assumption, that one can relax the integrality constraint on the purchase quantities, which is reasonable in cases in which the quantities are not very small (such as the case of cloud infrastructure capacity expansions).

For example, suppose that the provider purchases Q_1 units of cluster 1, whose parameters are a_{11} and a_{21}, together with Q_2 units of cluster 2, whose parameters are a_{12} and a_{22}. Alternatively, the provider can purchase Q'_1 units of cluster $1'$, whose parameters are a'_{11} and a'_{21}, together with Q'_2 units of cluster $2'$, whose parameters are a'_{12} and a'_{22}. The fixed expansion costs of the two expansion plans are the same. The holding costs of the two expansion plans are also the same provided that $a_{11}Q_1 + a_{12}Q_2 = a'_{11}Q'_1 + a'_{12}Q'_2$ and $a_{21}Q_1 + a_{22}Q_2 = a'_{21}Q'_1 + a'_{22}Q'_2$. With the integrality constraint relaxed, for any (Q_1, Q_2), there must exist a solution (Q'_1, Q'_2) to the two equations. Therefore, the problem boils down to a comparison of the total replenishment costs of the two expansion plans.

3.3.2.3 The performance gap between the two-cluster and the multi-cluster policies.

In this subsection, we provide a lower bound of the sum of the holding and the fixed expansion costs. If one can employ more cluster-types, the total cost decreases, but it can never be smaller than the lower bound. Hence, the difference between the total cost of employing two cluster-types and the lower bound gives the maximum value of employing additional cluster-types.

The lower bound of the total cost results from an ideal policy, for which the provider can re-customize the cluster configurations at every expansion point such that the supply can perfectly match the demand for both attributes until the next expansion point. In reality, however, the provider does not have such flexibility to re-customize at every expansion point, for various reasons described in the introduction, and thus the provider has to employ multiple (at least two) pre-configured cluster-types for an extended period of time (i.e., one CEC in our model). Clearly, as the number of the pre-configured cluster-types employed approaches infinity, the resulting total cost will approach (from above) that of the ideal policy, which is the lower bound.

Specifically, let TC^* denote the lower bound of the total cost. To achieve TC^*, the provider needs to employ a customized cluster-type at each (given or optimized) expansion point, $\tau_{k,n}$, whose parameters are $a^*_{1,k,n}$ and $a^*_{2,k,n}$. Let $Q^*_{k,n}$ be the purchase quantity of such a cluster-type at $\tau_{k,n}$.

$$a^*_{1,k,n}Q^*_{k,n} = D_1(\tau_{k,n}, \tau_{k+1,n}), a^*_{2,k,n}Q^*_{k,n} = D_2(\tau_{k,n}, \tau_{k+1,n}). \tag{3.21}$$

Next, consider that the provider adopts a simultaneous-replenishment policy with two pre-configure cluster-types, whose configurations are (a_{11}, a_{21}) and (a_{12}, a_{22}). To achieve a balance between supply and demand at each expansion time with those cluster-types, we

should have

$$a_{11}Q_{1,k,n} + a_{12}Q_{2,k,n} = D_1(\tau_{k,n}, \tau_{k+1,n}), a_{21}Q_{1,k,n} + a_{22}Q_{2,k,n} = D_2(\tau_{k,n}, \tau_{k+1,n}). \quad (3.22)$$

From (3.21) and (3.22), the holding costs of the two policies are equal. Moreover, for each replenishment, the fixed expansion cost of the simultaneous-replenishment policy with two clusters is βA, where $\beta \in [1, 2]$, while that of the ideal policy is only A. Thus, the optimal total cost of simultaneously employing two cluster-types can reach the lower bound, TC^*, if and only if $\beta = 1$.

Let $TC^{(2)}_{sim}(\beta)$ and $TC^{(2)}_{seq}$ denote the optimal total cost of the two-cluster model under the simultaneous- and the sequential-replenishment policies, respectively. According to Proposition 3.1, the optimal total cost of the two-cluster model, $TC^{(2)}$, can be written as follows:

$$TC^{(2)} = TC^{(2)}_{sim}(\beta), \text{ if } \beta \leq \overline{\beta}; \text{otherwise, } TC^{(2)} = TC^{(2)}_{sim}(\overline{\beta}) = TC^{(2)}_{seq}. \quad (3.23)$$

Therefore, at the extreme point where $\beta = 1$, the two-cluster model can achieve the lower bound, TC^*, and thus there is no need to employ more cluster-types. Moreover, as β increases, the gap between the total cost of the two-cluster model and the lower bound increases, but the gap is bounded by $\Delta^{(2)}$, which is equal to $TC^{(2)}_{sim}(\overline{\beta}) - TC^* = TC^{(2)}_{seq} - TC^*$.

The above analysis can shed some light on when it is good enough to employ only two cluster-types. (1) When the savings on the fixed expansion costs due to the joint replenishments are significant (i.e., when β is close to one), the total cost of employing two clusters (simultaneously) must be good enough. (2) Otherwise, the provider may prefer the sequential-replenishment policy to the simultaneous-replenishment policy, and the maximum value of employing more than two cluster-types is bounded by $\Delta^{(2)}$. Under conditions that lead to a relatively small $\Delta^{(2)}$, the two-cluster policy must be good enough; in Section 3.4.3,

we will explore such conditions numerically.

3.3.3 Analysis of the Replenishment Policies that Face Multiple Demand Scenarios

Our interviews with management of a leading cloud provider indicate that their forecasting team mainly provides the capacity team with point estimates of demand for each capacity attribute over time. Those point estimates consist of a projected growth in demand for that attribute, which can be used as the input to our base model. That said, because the estimates can never be accurate, the forecasting team sometimes also provides different demand scenarios and the associated probabilities of those scenarios taking place. In Microsoft (2010)'s report, it is stated: "The difficulty of predicting future need for computing resources and the long leadtime for bringing capacity online is another source of low utilization [of cloud servers]", which leads to overprovisioning. This implies Microsoft chooses the conservative demand scenarios so that it does not face demand shortage. In this section, we discuss how to extend our base model to a scenario analysis based on multiple demand scenarios.

Specifically, let $s \in \{1, 2, ..., S\}$ be the s-th demand scenario. Each demand scenario consists of two estimated demand growth functions, $D_i^{(s)}(a, b)$ (for any $a, b \in [0, T]$), for the two attributes $i = 1$ and 2. For instance, if the demand for attribute i under the s-th scenario is exponential with a rate $\lambda_i^{(s)}$, then that demand scenario can be characterized by $(\lambda_1^{(s)}, \lambda_2^{(s)})$. Let $p(s)$ be the probability of the s-th demand scenario taking place, then $\sum_{s=1}^{S} p(s) = 1$.

The cloud provider can derive the optimal capacity-expansion plan for a target demand scenario (e.g., the s-th scenario), using our base model in Section 3.2. Then, the provider can evaluate the actual total cost of implementing that specific capacity-expansion plan in the event of other demand scenarios (e.g., the s'-th scenario) taking place. Further, taking the probabilities of all demand scenarios into account, the cloud provider can obtain the expected total cost of implementing the specific expansion plan based on the target demand

scenario. Let $TC(s'|s)$ denote the actual total cost of implementing the capacity-expansion plan based on the s-th scenario, while the s'-th scenario is occurring. Thus, the expected total cost can be written as $EC(s) = \sum_{s'=1}^{S} p(s')TC(s'|s)$.

We would like to point out that, for the given expansion plan, the evaluation of the actual total cost faced by any given demand scenario is straightforward. For example, in today's cloud business environment, ensuring no capacity shortage is of paramount importance for the cloud infrastructure providers. As noted, cloud providers all target an extremely high service level (typically above 99%), as stated in their SLAs. Thus, it is most likely that the capacity-planning team of a cloud provider will target the scenario with the highest demand growth rates; for instance, $\lambda_i^{(s)} > \lambda_i^{(s')}$ for any $s' \neq s$ and $i = 1, 2$, if the demand functions are exponential. In that case, the cloud provider should expect extra holding costs in case that other (i.e., lower) demand scenarios are occurring, and the total extra holding cost is easy to calculate. That is, $TC(s'|s) = TC(s|s) + \sum_{i=1}^{2} h_i \int_0^T \left[D_i^{(s)}(0,t) - D_i^{(s')}(0,t) \right] dt$.

Although the common practice is to target the scenario with the highest demand growth rates, the cloud provider can consider choosing the best scenario to minimize the expected total cost $EC(s)$. This goal must be easy to achieve, given that $TC(s'|s)$ can be easily evaluated. The only issue is that, when the targeted demand scenario is not the one with the highest demand growth rates, there will be capacity shortages at times, and one needs to be mindful of having appropriate assumptions about the consequences of the shortages.

For the rest of this subsection, we allow for shortages. Then, let $\{d_i^{(1)}(t), d_i^{(2)}(t), ..., d_i^{(s)}(t), ..., d_i^{(S)}(t)\}$ be the set of all demand growth scenarios provided by the forecasting team, with $p^{(s)}$ being the probability of the s-th demand scenario taking place. Without loss of generality, we assume that the demand scenarios are ranked in an increasing order, $d_i^{(1)}(t) < d_i^{(2)}(t) < ... < d_i^{(s)}(t) < ... < d_i^{(S)}(t)\}$ for all t. Our goal is to calculate the expected total cost of employing the capacity expansion plan, which is developed based on a specific demand scenario

chosen by the provider. Suppose that the provider decides to make the capacity expansion plan based on the s-th demand scenatio, where $s \in \{1, 2, ..., S\}$. By choosing this demand scenario, the set of other demand scenarios can be categorized into two subsets as follows:

$$\Gamma_i^{(H)} = \{s' | d_i^{(s')} > d_i^{(s)}; t \in [0, T]\} \text{ and } \Gamma_i^{(L)} = \{s' | d_i^{(s')} < d_i^{(s)}; t \in [0, T]\}.$$

$\Gamma_i^{(H)}$ (or $\Gamma_i^{(L)}$) contains scenarios with higher (or lower) forecasted demand growths than the chosen scenario. For demand scenarios in $\Gamma_i^{(L)}$, the provider faces extra holding costs compared to the chosen scenario. Specifically, the extra holding cost of a demand scenario $s' \in \Gamma_i^{(L)}$ is

$$h_i \int_0^T [D_i^{(s)}(0, t) - D_i^{(s')}(0, t)]dt, \ i = 1, 2, \forall s' \in \Gamma_i^{(L)} \tag{3.24}$$

For demand scenarios in $\Gamma_i^{(H)}$, however, the provider faces lower holding costs while incurs shortage costs compared to the chosen scenario. We assume that the unmet demand will be lost due to the competitive cloud market; that is, unmet demand will be attracted to other providers.

Here, we present the result based on the m-EI-sequential policy only. Under the sequential-replenishment policy, we again focus on analyzing the situation where cluster 1 is the leading cluster since the opposite situation can be analyzed in the same vein. Define $\tau_{k,n} + x_{k,n}$ as the time when the excess capacity of attribute 2 reaches zero after the k-th replenishment in the n-th cycle. Note that $x_{k,n} < t_n = \tau_{k+1,n} - \tau_{k,n}$ is uniquely determined as follows:

$$D_2^{(s)}(\tau_{k,n}, \tau_{k+1,n}) = D_2^{(s')}(\tau_{k,n}, \tau_{k,n} + x_{k,n}), \forall x' \in \Gamma_i^{(H)}.$$

Recall that $t_n (n = 1, ..., m)$ denote the constant intervals between consecutive replenishments of the leading cluster under either the EI-sequential or the m-EI-sequential policies. Let

$\pi_i (i = 1, 2)$ be the shortage cost per unit of attribute i due to the lost sales. Moreover, define

$$\begin{aligned} F_1(n,k) &= t_n \left[\frac{a_{11}}{a_{21}} D_2^{(s)}(T_n, \tau_{k+1,n}) - \sum_{l=1}^{k-1} D_1^{(s')}(\tau_{l,n}, \tau_{l,n} + x_{l,n}) \right] \\ &\quad - \int_{\tau_{k,n}}^{\tau_{k,n}+x_{k,n}} D_1^{(s')}(\tau_{k,n}, t)dt - D_1^{(s')}(\tau_{k,n}, \tau_{k,n} + x_{k,n})(t_n - x_{k,n}) \end{aligned}$$

$$\begin{aligned} F_2(n) &= (T_{n+1} - \tau_{m_n+1,n}) \left[\frac{a_{12}}{a_{22}} D_2^{(s)}(\tau_{m_n+1,n}, T_{n+1}) + \frac{a_{11}}{a_{21}} D_2^{(s)}(T_n, \tau_{m_n+1,n}) - \sum_{l=1}^{m_n} D_1^{(s')}(\tau_{l,n}, \tau_{l,n} + x_{l,n}) \right] \\ &\quad - \int_{\tau_{m_n+1,n}}^{\tau_{m_n+1,n}+x_{m_n+1,n}} D_1^{(s')}(\tau_{m_n+1,n}, t)dt - D_1^{(s')}(\tau_{m_n+1,n}, \tau_{m_n+1,n} + x_{m_n+1,n})(T_{n+1} - \tau_{m_n+1,n} - x_{m_n+1,n}) \end{aligned}$$

The next lemma gives the holding and shortage costs, if the provider adopts a sequential replenishment policy based on the s-th scenario, while the s'-th demand scenario taking place.

Lemma 3.4. *Facing the s'-th demand scenario, where $s' \in \Gamma_i^{(H)}$, employing the capacity expansion plan based on the s-th scenario leads to the following costs:*

- *Holding costs:*

$$HC_1^{(s')} = h_1 \sum_{n=1}^{N} \left[F_2(n) + \sum_{k=1}^{m_n} F_1(n,k) \right] \tag{3.25}$$

$$HC_2^{(s')} = h_2 \sum_{n=1}^{N} \left[\sum_{k=1}^{m_n+1} \left(x_{k,n} D_2^{(s)} - \int_{\tau_{k,n}}^{\tau_{k,n}+x_{k,n}} D_2^{(s')}(\tau_{k,n}, t)dt \right) \right] \tag{3.26}$$

- *Shortage cost:*

$$SC_1^{(s')} = \pi_1 \sum_{n=1}^{N} \left[\sum_{k=1}^{m_n+1} D_1^{(s')}(\tau_{k,n} + x_{k,n}, \tau_{k+1,n}) \right] \tag{3.27}$$

$$SC_2^{(s')} = \pi_2 \sum_{n=1}^{N} \left[\sum_{k=1}^{m_n+1} D_2^{(s')}(\tau_{k,n} + x_{k,n}, \tau_{k+1,n}) \right] \tag{3.28}$$

Define $\Delta HC_i^{(s')} = HC_i^{(s')} - HC_i^{(s)}$ for any $s' \in \Gamma_i^{(H)}$.

Proposition 3.7. *The total cost of scenario s' is*

- *If $s' \in \Gamma_i^{(L)}$, then $TC^{(s')} = TC^{(s)} + \sum_{i=1}^{2} \left[h_i \int_0^T [D_i^{(s)}(0,T) - D_i^{(s')}(0,T)]dt \right]$.*
- *If $s' \in \Gamma_i^{(H)}$, then $TC^{(s')} = TC^{(s)} + \sum_{i=1}^{2} \Delta HC_i^{(s')} + \sum_{i=1}^{2} SC_i^{(s')}$.*

Finally the expected cost of implementing the capacity plan based on the s-th scenario is $\sum_{s'=1}^{S} p^{(s')} TC^{(s')}$. To optiamlly choose a demand scenario, based on which the capacity expansion will be developed, the provider needs to calculate the expected cost of implementing the capacity plan based on every scenario and choose the one leading to the lowest expected cost.

In summary, our base model can be readily extended to the scenario analysis outlined at the beginning of this section. The convenience of the extension is due to the fact that, for any given capacity-expansion plan (derived from the chosen demand scenario), the numerical evaluation of the actual total cost faced by other demand scenarios is straightforward. Then, one can easily obtain the expected total cost of implementing that given capacity-expansion plan. Alternatively, one can request the forecasting team to estimate the demand distributions for all time points, instead of just the demand scenarios, and study a model, using the estimated demand distributions. It is worth noting, however, that our base model and the scenario analysis outlined in this section are valuable as they are very much in line with

the most common practices at the leading cloud provider that we interviewed. The study of the alternative model is beyond the scope of this dissertation; thus, we leave it for future research.

3.4 Numerical Study

In this section, we present our numerical study to gain more insight into the performance of the policies proposed in the previous sections. Specifically, in Section 3.4.1, we compare the three practical policies of Section 3.2, namely, the EI-Sequential, the m-EI-Sequential, and the EI-Simultaneous policies. In Section 3.4.2, we compare the two algorithms of Section 3.3.1, the DP-algorithm and the FL-heuristic, which aim to find the optimal number and lengths of the CECs. Finally, in Section 3.4.3, we provide the performance of the cluster-selection heuristic proposed in Section 3.3.2 and explore important conditions, under which employing only two cluster-types is good enough.

In our numerical experiments, we set the planning horizon, T, to either 6 or 12 months. Demand for new capacity of attribute i is $D_i(a,b) = \gamma(e^{\lambda_i b} - e^{\lambda_i a})$. We set $\gamma = 10^5$ because such a scale is close to the number of servers in a cloud provider's data centers. For example, Google is estimated to have more than 900,000 servers in its data centers worldwide (Nimrodi 2014). We set $\lambda_1 = 0.05$, resulting in an annual growth rate of around 80%, and let $\lambda_2 \in \{0.25, 0.50, 0.75\} \times \lambda_1$. The range of these growth rates is wide enough to cover the actual demand growth rates of the leading cloud providers[4].

For the holding cost rates, we set $h_1 = 1$ and let $h_2 \in \{0.5, 1.0, 2.0\} \times h_1$. For the fixed expansion cost, we examined $A \in \{0.050, 0.075\} \times \gamma$. Different cluster-types are characterized by their a_{ij}'s. We set $a_{11} = 50$ and $a_{12} = 10$ and define $R_j = a_{1j}/a_{2j}$ as the ratio of attributes

[4]There is evidence about the revenue growth rates of some cloud providers, which is an indicator of the demand growth rates. For AWS, the revenue growth rates were around 45% in recent years (https://ir.aboutamazon.com/quarterly-results), whereas for Microsoft Azure, the revenue growth rate reached 89% in 2018 (Protalinski 2018).

of cluster $j \in \{1, 2\}$. Note that given the fixed a_{11} and a_{12}, varying a_{21} and a_{22} is equivalent to varying R_1 and R_2, respectively. In Sections 3.4.1 and 3.4.2, we present our examination of $R_1 \in \{6, 8, 10\}$ and $R_2 \in \{0.4, 0.6, 0.8\}$; thus, in total, there are 162 instances in each of the two sections. In Section 6.3, which focuses on the cluster-selection problem, we present an expanded test bed with even more possibilities of R_1 and R_2. The collection of R_1 and R_2 in our numerical study includes various ratios of attributes that are close to the typical ratios of the numbers of CPU to RAM in reality[5].

Finally, it is worth noting that the entire setting of our numerical study ensures that the resulted frequency of replenishments resembles the possible frequency that a cloud provider purchases and employs the servers in practice (e.g., monthly, quarterly).

3.4.1 The Sequential- and Simultaneous-Replenishment Policies

In the comparison between the sequential- and simultaneous-replenishment policies, we implement the analytical results of Sections 1.4.1 and 1.4.2, respectively, to obtain the optimal policy parameters, m^*, as well as the total costs of those policies. In this section, $T = 6$ months.

Specifically, under a sequential-replenishment policy (the EI-Sequential or the m-EI-Sequential policy), we consider two cases, where cluster-type $l \in \{1, 2\}$ is the leading cluster. Then, we calculate the total cost of the two cases and choose the one that results in the smaller cost. Here, we obtain m^* from a numerical search for the optimal integer value that maximizes the exact total cost function (3.12). Similarly, under the simultaneous-replenishment policy (i.e., the EI-Simultaneous policy), we obtain m^* from a search for the optimal integer value that maximizes (3.17).

---

[5]There is information about the configurations of the virtual machines. For example, at Google Compute Engine, the high-RAM machines have 6.5 GB of memory per vCPU, while the high-CPU machines have 0.9 GB of memory per vCPU (https://cloud.google.com/compute/docs/machine-types).

First, based on the 162 instances tested, the increased cost of employing the EI-Sequential policy is, on average, 17.37% of the total cost of employing the m-EI-Sequential policy. Note that, similar to the EI-Sequential policy, the m-EI-Sequential policy has equal replenishment intervals, with the exception of only the last interval. Hence, the significant cost savings of the m-EI-Sequential policy can justify having flexibility in determining the last replenishment time.

Second, our numerical study confirms the findings of Proposition 3.1. There is always a unique threshold of β, above which the m-EI-Sequential policy dominates the EI-Simultaneous policy. To provide an example, Figure 3.4 depicts the total costs of those policies at different values of $\beta \in [1, 2]$. In this example, the total cost of the EI-Simultaneous policy is smaller than that of the m-EI-Sequential policy if and only if β is smaller than a threshold, which is between 1.2 and 1.3.

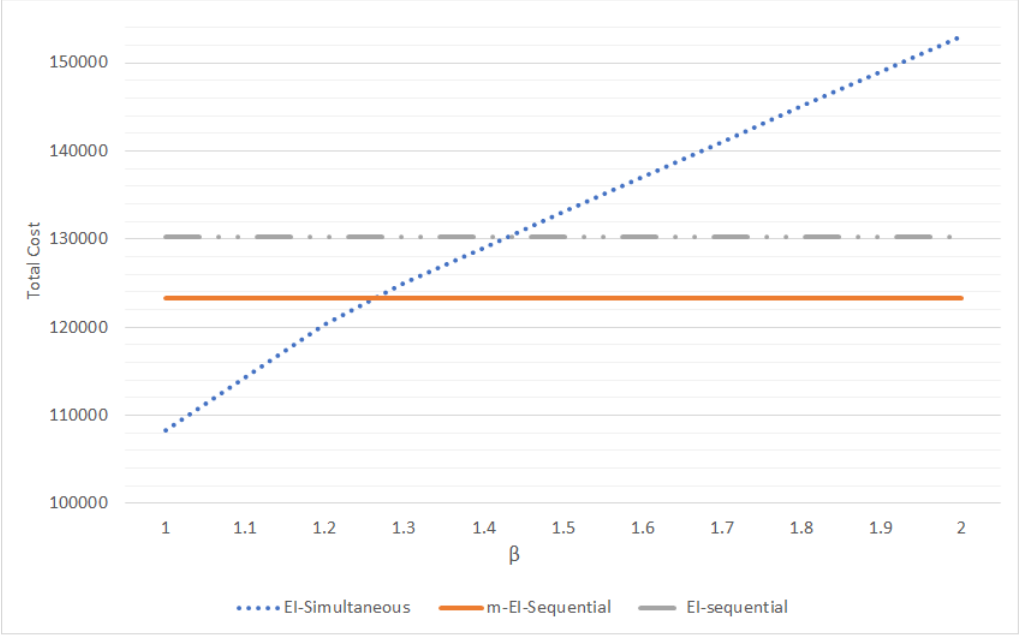


Figure 3.4: The sequential- and simultaneous-replenishment policies. $T = 6$, $\gamma = 10^5$, $\lambda_1 = 0.05$, $\lambda_2 = 0.75 \times \lambda_1$, $h_1 = 1$, $h_2 = 0.5$, $R_1 = 10$, and $R_2 = 0.8$.

Finally, we have observed that if $h_i (i \in \{1, 2\})$ increases or A decreases, then m^* increases as expected. We highlight the more remarkable results in Observations 1 to 4.

Observation 3.1. *As $\lambda_i (i \in \{1, 2\})$ increases, m^* of the EI-Simultaneous policy weakly increases, whereas, as λ_f/λ_l increases, m^* of the m-EI-Sequential policy decreases.*

Consider a representative case whereby $A = 5000$, $h_2 = h_1$, $R_1 = 6$, and $R_2 = 0.4$. As λ_2 increases from 0.0125 to 0.0375, m^* of the EI-Simultaneous policy *increases* from 4 to 5. In contrast, given that cluster 1 is chosen as the leading cluster, as $\lambda_f/\lambda_l = \lambda_2/\lambda_1$ increases from 0.25 to 0.75, m^* of the m-EI-Sequential policy *decreases* from 4 to 2.

The explanation for Observation 3.1 is as follows: The simultaneous-replenishment policy requires matching supply with demand for both attributes at every expansion point. Consequently, a more rapid demand growth of any attribute leads to more frequent replenishments. The sequential-replenishment policy, by contrast, does not have such a requirement at expansion points when the leading cluster is employed. For instance, if cluster 1 is the leading cluster, then the excess capacity of attribute 1 will accumulate until the last expansion time. Only at the last expansion time, when the following cluster is employed, does the policy require matching supply with demand such that capacities of both attributes are leveled at the end of the cycle. As a result, when λ_f/λ_l increases, the last expansion interval, $T_{n+1} - \tau_{m+1,n}$, will increase, which can be seen in (3.8). Because $T_{n+1} - \tau_{m+1,n}$ will increase, $\tau_{m+1,n} - T_n$ will decrease, which in turn leads to the decrease in m^*.

3.4.2 The DP-Algorithm and the FL-Heuristic

We now study the optimal division of the planning horizon into multiple CECs. In this section, we consider a relatively long planning horizon (i.e., $T = 12$ months) compared to that of the previous section (i.e., $T = 6$ months), as we allow the planning horizon to consist of multiple CECs. We first implement the DP-algorithm to obtain the optimal cycle lengths

and then examine the performance of the FL-heuristic. Here, we assume that the optimal m-EI-Sequential policy is employed within each cycle. We define the *relative increased cost of the FL-heuristic* as the ratio (percentage) of the increased cost of using the heuristic solution to the total cost of using the solution of the DP-algorithm. Table 3.3 presents the statistics of the relative increased cost based on the 162 instances.

Table 3.3: Relative increased cost of the FL-heuristic compared to the DP-algorithm.

Minimum	1st Quartile	2nd Quartile	3rd Quartile	Maximum
0.00%	0.00%	0.36%	1.60%	12.27%

In general, the proposed FL-heuristic performs very well. The average of the relative increased cost is 1.63%, and the median is 0.36%. As noted in Section 3.3.1, the satisfactory performance of the heuristic makes its solution an excellent starting point of the DP-algorithm to reduce computational time. Moreover, the heuristic has distinct merits of implementation: (i) it is easy to communicate in practice; and (ii) its forward-looking nature makes it easy to deal with the rolling horizon effect.

However, we also should be mindful of a few cases in which the fixed expansion cost and the discrepancy between λ_1 and λ_2 are large (we observed two cases with relative increased above 10%). In such cases, the heuristic suggests more frequent expansions than the DP-algorithm does, causing increased costs that cannot be ignored. We attribute the increased costs in those cases to the myopic nature of the heuristic. Note that at the beginning of each cycle, the heuristic aims to minimize the average cost rate of that cycle only, without considering the large expansion costs and the increasing discrepancy between the demand growth rates in the later cycles, which leads to a significant deviation from the optimal cycle length.

Moreover, we make two observations about the optimal solution of the DP-algorithm:

Observation 3.2. *For attributes i and j whereby $\lambda_i < \lambda_j$, as the gap between λ_i and λ_j increases (i.e., λ_i/λ_j decreases), the optimal number of CECs usually decreases. Moreover, the cycle lengths usually decrease as we move toward the end of the planning horizon.*

For example, when $A = 5000$, $h_2 = 0.5$, $R_1 = 10$, and $R_2 = 0.8$, we find: (i) when $\lambda_2 = 0.25\lambda_1$, there are two CECs, whose lengths are 8 and 4; (ii) when $\lambda_2 = 0.50\lambda_1$, there are four CECs, whose lengths are 4, 3, 3, and 2; and (iii) when $\lambda_2 = 0.75\lambda_1$, there are six CECs with equal lengths of 2.

To explain the first part of the above observation, note that as the discrepancy between λ_i and λ_j enlarges, the growths of demand for the two attributes become more disproportionate, which makes it more difficult for the provider to balance supply with demand for both attributes, using the pre-configured cluster-types with fixed ratios of attributes. Nevertheless, at the end of every CEC, the capacities of both attributes are required to be leveled (i.e., one must achieve a balance between supply and demand for both attributes). Hence, as the discrepancy between λ_i and λ_j increases, it is more difficult to achieve the required balance at the end of every CEC, and, thus, it is desirable to have fewer CECs. To explain the second part of the above observation, note that the demand growths become more rapid toward the end of the planning horizon, as the demand functions are exponential. To satisfy the more rapid demand growths, the provider should have shorter cycle lengths and more frequent expansions toward the end of the planning horizon.

Observation 3.3. *The total cost usually increases if any of the two ratios of attributes of the leading and the following clusters, $R_l = a_{ll}/a_{fl}$ and $R_f = a_{lf}/a_{ff}$, increases.*

Consider an example whereby $A = 7500$, $\lambda_2 = 0.025$, $h_2 = 2$, and cluster 1 is the leading cluster. When $R_l = R_1 = 8$ fixed and as $R_f = R_2$ increases from 0.4 to 0.8, the total cost increases by 13.1%; when $R_f = R_2 = 0.4$ fixed and as $R_l = R_1$ increases from 6 to 10, the total cost increases by 33.1%.

To explain this observation, we note that, as R_l ($l \in \{1, 2\}$) increases, the excess capacity (of attribute l) will increase when the leading cluster is employed, and as R_f increases, the excess capacity will be depleted more slowly when the following cluster is employed. As a result, the total cost will increase due to the increased total holding cost of the excess capacity. This observation also supports the rationale for the cluster-selection heuristic, which prefers smaller R_l and R_f.

3.4.3 The Cluster-Selection Problem

The objective of this section is twofold: First, we examine the performance of the proposed cluster-selection heuristic for the sequential-replenishment policy with two cluster-types. Second, we find critical conditions under which the total cost of employing only two cluster-types is close to the lower bound provided in Section 3.3.2. The important implication is that, under those conditions, employing only two cluster-types must be as good as employing more than two cluster-types.

Because the ratio of the demand growth rates, the ratio of the holding cost rates, and the fixed expansion cost all have considerable influence on the optimal selection to the cluster-selection problem, we have expanded the test bed of these parameters. That is, $\lambda_2 \in \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \frac{6}{7}, \frac{7}{8}\} \times \lambda_1$ and $h_2 \in \{\frac{1}{6}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 6\} \times h_1$, and $A \in \{0.050, 0.075, 0.100, 0.150, 0.200\} \times \gamma$. We also have expanded the test bed of the cluster configurations: $R_1 \in \{6, 7, 8, ...15\}$ and $R_2 \in \{0.1, 0.2, ..., 0.8\}$. In total, there are 315 instances, and for each instance, there are 80 possible pairs of cluster-types.

We focus on one CEC with $T = 6$ months, we consider the optimal selection of two cluster-types at the beginning of the cycle, and we adopt the sequential-replenishment policy because the problem under the simultaneous-replenishment policy is straightforward, as noted in Section 5.3.2. To determine the optimal pair of cluster-types, we first use the exhaustive

search method. Then, we implement the cluster-selection heuristic and calculate the total cost of using the heuristic solution.

We report the relative increased cost of the cluster-selection heuristic compared to the minimal cost of using the optimal pair of cluster-types. Table 3.4 provides the summary statistics.

Table 3.4: The relative increased cost of using the cluster-selection heuristic compared to the minimal cost.

First Quartile	Second Quartile	Third Quartile	Maximum	Average
0.00%	0.09%	0.74%	2.97%	0.48%

As Table 3.4 shows, the cluster-selection heuristic performs very well in selecting *two* cluster-types. The more important question is: *When is the employment of only two cluster-types good enough*?

As noted in Section 3.3.2.3, if the provider adopts the sequential-replenishment policy, the cost savings of employing more than two cluster-types will never exceed $\Delta^{(2)} = TC^{(2)}_{seq} - TC^*$. Here, we define $\Delta^{(2)}/TC^*$ as the maximum benefit (in percentage) of employing more than two cluster-types under the sequential-replenishment policy, and we examine when such a benefit is relatively small.

Observation 3.4. *The maximum benefit of employing more than two cluster-types under the sequential-replenishment policy is relatively small if (i) h_l is smaller than h_f , or (ii) λ_l/λ_f is large.*

To provide an example, Figure 3.5 depicts the maximum benefit of employing more than two cluster-types under the sequential replenishment policy. Note that, under the particular setting of this example, the optimal pair of clusters is selected, whereby cluster 2 is optimally

chosen as the leading cluster, i.e., $l = 2$ and $f = 1$. On the left graph, as $h_l = h_2$ decreases, such a benefit decreases; on the right graph, as $\lambda_l/\lambda_f = \lambda_2/\lambda_1$ increases, such a benefit also decreases.

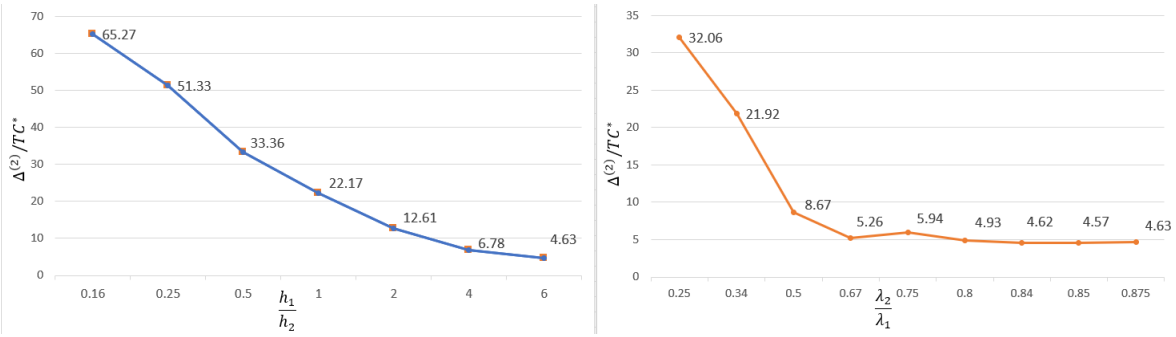


Figure 3.5: The maximum benefit of employing more than two types under the sequential-replenishment policy. $T = 6$, $\gamma = 10^5$. Left: $\lambda_2 = (7/8)\lambda_1$ is fixed; Right: $h_2 = (1/6)h_1$ is fixed. For both graphs, cluster 2 is optimally chosen as the leading cluster, i.e., $l = 2$ and $f = 1$, and the optimal pair of clusters consists of $R_1 = 15$ and $R_2 = 0.8$.

The reason is that, when employing the sequential-replenishment policy with two cluster-types, the excess capacity of attribute l will accumulate with the employment of the leading cluster (i.e., cluster l, where $l \in \{1, 2\}$). Thus, any condition that leads to a lower holding cost rate (i.e., condition (i) above) or less excess capacity of attribute l (i.e., condition (ii) above) must be a favorable condition for that policy. Under those favorable conditions, the total cost of employing the sequential-replenishment policy with two cluster-types can be close to that of employing the ideal policy, and thus the potential benefit of employing additional cluster-types is small.

In summary, Section 3.3.2.3 and Observation 3.4 together provide guidelines for the employment of only two, rather than many, cluster-types, echoed by the cloud providers' desire to reduce the variety of cluster-types in practice. That is, when the synergies in the fixed expansion costs are significant, the provider should adopt the simultaneous-replenishment

policy, and, as a result, using two cluster-types can be almost as effective as using many cluster-types. In contrast, when the synergies in the fixed expansion costs are negligible, the provider should adopt the sequential-replenishment policy, and, as a result, the employment of only two cluster-types may result in a significant loss of efficiency; if the two conditions of Observation 3.4 are met, however, the maximum benefit of employing additional cluster-types is small, which implies that the employment of only two cluster-types can be almost as effective as the employment of many cluster-types.

3.5 Conclusion

In this chapter, we focused on cloud industry. On the one hand, the key determinants of configurations of physical and virtual machines, such as CPU and RAM, are often added in pre-configured server clusters, instead of individual capacity scaling. That is, supply of RAM and CPU is bundled in servers. On the other hand, demand for RAM and CPU is time-dependent and disproportionate to supply ratio. Therefore, the optimal capacity expansion policy, in this environment, has become a critical challenge. To tackle this problem, the planning horizon is divided into multiple CECs, and a family of capacity expansion policies is analyzed. In each cycle, capacities of the two attributes are expanded through acquisitions, either sequentially or simultaneously, of the two pre-configured cluster-types. The goal is to match supply and demand at the end of each cycle. Within each cycle, we derived the optimal timing and quantities of the expansions. Next, to find the optimal timing of the cycles, we devised a DP-based algorithm and a forward-looking heuristic. Finally, we investigated the optimal selection of two cluster-types from many possible configurations. To that end, on top of the exhaustive search approach, we proposed a cluster-selection heuristic. The cluster-selection heuristic is as follows: Choose the configuration of the leading cluster that will lead to the minimal excess capacity when it is employed consecutively, and then

choose the configuration of the following cluster that will lead to the fastest depletion of the excess capacity when it is employed in the end. Applying our proposed policy, a cloud company is able to better match supply and demand of CPU and RAM, while minimizing the total cost.

Chapter 4

THE IMPACT OF 3D PRINTING ON MANUFACTURER-RETAILER SUPPLY CHAINS

3D printing, also known as additive manufacturing, is an alternative manufacturing technique that is based on producing a product layer by layer. This technique contrasts with traditional manufacturing techniques, such as milling, forging, and welding. 3D printing has received significant and growing attention in recent years. Perhaps the most prominent mention of 3D printing was during President Obama's 2013 State of the Union address (Gross , 2013), where the President said *"A once-shuttered warehouse is now a state-of-the art lab where new workers are mastering the 3D printing that has the potential to revolutionize the way we make almost everything."* President Obama was referring to the National Additive Manufacturing Innovation Institute (NAMII) in Youngstown, Ohio, created in part due to a $30 million government investment, whose objective is to revolutionize manufacturing using 3D printing[1] .

Conner et.al (2014) propose a framework that attempts to understand where 3D printing is economically feasible. They explain that the key attributes of a product are its complexity, customizability, and production volume; they unify these three characteristics into a Modified Complexity Factor (MCF). They argue that 3D printing is likely to have a lower unit production cost than traditional manufacturing techniques if the MCF is high, which occurs with high complexity, and/or high customizability, and/or low production volume. Complex products, in particular, can have designs that simultaneously reduce material while increasing

[1] www.3ders.org

strength (e.g., a lattice), which are very difficult to manufacture using traditional techniques, but pose little problem for 3D printing. A variety of real products (ranging from artwork, automotive and aviation parts, dies, dental braces and even smartphones) is discussed in Conner et.al (2014); the references therein provide further examples. A notable example is a ball bearing assembly that would require 18 parts being manufactured separately using traditional techniques, which would then need to be subsequently assembled; 3D printing, in contrast, can manufacture/assemble them simultaneously. This chapter examines both types of products with higher and lower 3D printing unit cost compared to traditional manufacturing unit production cost, and we discuss under what economic conditions 3D printing is preferred to traditional manufacturing techniques.

In addition to the United States, other governments are investing in 3D printing. In 2014, the Japanese Economy, Trade, and Industry Ministry (METI) selected several universities and technical schools to receive a $44 million subsidy for bringing 3D printing technology into their institutions[2], where the goal was to introduce 3D printing technology into small and medium-sized enterprises through university students. In 2015, the Chinese Ministry of Science and Technology announced a project with a 2 billion RMB (approximately $313 million) investment in 3D printing technology over three years[3]. The Indian government supports local manufacturing through a Make-in-India program[4] , where local production of goods is facilitated through 3D printing. Hall (2016) reports that the French government, through the FAIR (French acronym for Additive Manufacturing for the Intensification of Reactors) collaborative project, will distribute €10.5 million to French companies for 3D printing. Two innovative European production institutions, Brightlands Chemelot Campus (in Netherlands) and Centexbel (in Belgium), and Katholieke Universiteit Leuven Kulak

[2]www.3ders.org

[3]http://www.3ders.org/articles/20151118-china-to-invest-300-million-in-3d-printing-rd.html

[4]http://www.itroadmap.in/feature/outlook-indian-3d-printing-market-2016

have entered into a cross-border partnership in the field of 3D printing[5]; the goal of this project is to create a 3D printing hub, and €1.3 million of its €3 million fund is provided by a subsidy from the Interreg program of the European Union.

Prominent manufacturers and manufacturing industries, that have historically utilized traditional manufacturing techniques, are adopting 3D printing technology. For example, General Electric is committed to a $3.5B-investment in 3D printing and aims to produce over 100,000 3D printed parts for its LEAP and GE9X engines by 2020[6]. Disney has also invested in 3D printing and in 2016 they filed for three new related patents (Grunewald , 2016). In healthcare, Grunewald (2016) reports that the 3D printing industry had total revenues of $487 million in 2014 with the expectation of 18.3% annual growth through 2020. There are even projections of 3D printing human organs[7]. Licata (2013) claims that 3D printing leads to more sustainable supply chains via 3D printed three-dimensional solar panels; this article also suggests that supporting 3D printing would help the United States to reach its 2020 carbon reduction goal.

Although this chapter is motivated by 3D printing technology, the models and results are an extension to Dual Sourcing literature. The results in Section 4.3 can be applied to the following setting. Suppose there exists a manufacturer which orders product X from an off-shore supplier with a long lead time, and she has the opportunity to invest in an on-shore production line with a short lead time to produce product X. The questions she faces are: 1) whether or not to invest in the new production line, and 2) if investing, what should the capacity of the new production line be? The setting and questions are similar to the scenario where the retailer opts for 3D printing, hence the results can be applied.

---

[5]https://www.brightlands.com/news-events/3d-printing-project-chemelot-campus-european-subsidy

[6]http://3dprintingreviews.blogspot.co.uk/2013/06/ge-aviation-to-grow-better-fuel-nozzles.html

[7]http://www.cnn.com/2014/04/03/tech/innovation/3-d-printing-human-organs

4.1 Benchmark Model without 3D Printing

In this section we present our benchmark model that captures the traditional situation where 3D printing is not available to firms. This model serves two purposes: 1) the constructs in this section are useful to present our results for the situation where firms have access to 3D printing, and 2) we may compare the firms' outcomes of the benchmark model with those under 3D printing.

Our benchmark model consists of a single manufacturer selling a single product to a single retailer via a wholesale-price contract with unit wholesale price $w > 0$, which is based on the model in Lariviere and Porteus (2001). The retailer, in turn, sells to (continuous) stochastic customer demand D with density f and distribution F, with support on $[0, U]$. The manufacturer's unit production cost is $c_m > 0$ and the retailer's unit selling price is $r > 0$; to avoid trivial solutions, we assume that $c_m \leq w \leq r$. The retailer's single decision, as a function of w, is to determine the order quantity $q(w) \geq 0$ that maximizes his expected profit

$$\pi_R^b(r, w) = \max_{q \geq 0} E[r \min\{q, D\}] - wq, \tag{4.1}$$

which is solved by the classic Newsvendor solution $q_{nv}(w) = F^{-1}\left(1 - \frac{w}{r}\right)$. The subscript R refers to the retailer (M will refer to the manufacturer) and the superscript b refers to the benchmark case in this section. We assume that the retailer participates and purchases the product if and only if his resulting profits are at least A, which represents the opportunity cost of alternative business opportunities.

The manufacturer, being the leader in a Stackelberg game, has a single decision of determining the appropriate wholesale price w to maximize her profits, which is determined by

solving

$$\begin{aligned} \pi_M^b = \quad & \max_{w \geq 0} \quad (w - c_m) q_{nv}(w) \\ & \text{s.t.} \quad \pi_R^b(w) \geq A, \end{aligned}$$

where the constraint is to make sure that the retailer's maximized profit is at least A, to ensure participation. In order to solve the manufacturer's problem cleanly, we make the standard assumption that the demand distribution F has an increasing generalized failure rate: $xf(x)/(1 - F(x))$ is increasing in x. Most common distributions (e.g., uniform, truncated normal, gamma, Pareto) satisfy this constraint; see Banciu and Mirchandani (2013) for a more complete list of distributions.

Assumption 4.1. *The demand distribution F has an increasing generalized failure rate.*

The solution to the manufacturer's problem for $A = 0$ is summarized in the following proposition.

Proposition 4.1. *(Lariviere and Porteus , 2001) The manufacturer's profit function $(w - c_m)q_{nv}(w)$ is strictly unimodal in w, with optimality condition*

$$(1 - F(q)) \left(1 - \frac{qf(q)}{1 - F(q)} \right) = \frac{c_m}{r}. \tag{4.2}$$

Let q_s denote the unique solution to this optimality condition and let $w_s = r(1 - F(q_s))$ denote the associated optimal wholesale price.

If $A > 0$, then it is useful to define the maximum wholesale price or, equivalently, the minimum order quantity that will induce the retailer to participate.

Definition 4.1. *Let $w_m^A = \max\left\{w \geq 0 : \pi_R^b(r, w) \geq A\right\}$ and let $q_m^A = q_{nv}(w_A)$ denote the associated order quantity.*

Proposition 4.1 and Definition 4.1 lead us to the equilibrium solution of the manufacturer-retailer Stackelberg game in our benchmark situation, which we summarize in the following corollary.

Corollary 4.1. *The equilibrium wholesale price is* $w_b = \min\{w_s, w_m^A\}$ *and the equilibrium order quantity is* $q_b = q_{nv}(w_b)$*, which results in retailer profit* $\pi_R^b(r, w_b)$ *and manufacturer profit* $\pi_M^b = (w_b - c_m)q_{nv}(w_b)$*. Furthermore, if* $w_m^A \leq w_s$*, then* $\pi_R^b(r, w_b) = A$*.*

Finally, we make the assumption that $w_b \geq c_m$, so that the manufacturer is guaranteed to earn non-negative profit.

4.2 Manufacturer May Purchase 3D Printers

In this section we study a situation where the manufacturer has the option of purchasing multiple 3D printers at a unit cost of $K > 0$ per printer, to supplement/replace her traditional manufacturing capabilities. Let $n \geq 0$ denote the number of 3D printers, which is a decision variable under the manufacturer's control. Each 3D printer has the capacity to produce $Q > 0$ units over a fixed production horizon, and the unit printing cost is $c_p > 0$. To adopt 3D printing, we assume the manufacturer must incur a fixed cost $S > 0$ to convert existing manufacturing plans into 3D schematics. As in our benchmark, traditional manufacturing has a unit cost c_m. We let $v \in \{0, 1\}$ denote the manufacturer's binary decision to purchase one or more 3D printers. The manufacturer also decides how many of the products to 3D print, which we denote as $q_p \geq 0$, and how many of the products to manufacture, which we denote as $q_m \geq 0$. The final manufacturer decision is the wholesale price $w \geq 0$, which influences the retailer's order quantity $q_{nv}(w) = F^{-1}\left(1 - \frac{w}{r}\right)$. The manufacturer's profit-

maximization problem is

$$\begin{aligned} \pi_M^{3D} = \max_{\substack{n,q_p,q_m,w\geq 0 \\ v\in\{0,1\}}} \quad & wq_{nv}(w) - c_p q_p - c_m q_m - (nK+S)v \\ \text{s.t.} \quad & q_p \leq nQ \\ & nQ \leq vq_{nv}(w) \\ & q_p + q_m = q_{nv}(w) \\ & \pi_R^b(r,w) \geq A. \end{aligned} \tag{4.3}$$

The first constraint ensures that the number of 3D printed products will not exceed the capacity of the 3D printers. The second constraint ensures that the capacity of the 3D printers does not exceed the retailer's order size if 3D printers are utilized ($v = 1$), and forces the number of 3D printers to zero ($n = 0$) otherwise. The third constraint requires that the total products produced, using both 3D printing and traditional manufacturing, equals the retailer's order size. The fourth constraint ensures retailer participation, since the retailer's problem is unchanged from the benchmark problem in Equation (4.1). It is useful to reapply the well-known result from Lariviere and Porteus (2001), using the total unit 3D printing cost $c_p + \frac{K}{Q}$, rather than c_m, which allows us to define a new optimal order quantity $\hat{q}_s$ and associated wholesale price $\hat{w}_s$.

Corollary 4.2. *(Lariviere and Porteus , 2001) If the manufacturer uses 3D printing to produce all the products (i.e. $q_p = q_{nv}(w)$ and $n = \frac{q_{nv}(w)}{Q}$), then the manufacturer's profit function $(w - c_p - \frac{K}{Q})q_{nv}(w) - S$ is strictly unimodal in w, with optimality condition*

$$(1 - F(q))\left(1 - \frac{qf(q)}{1-F(q)}\right) = \frac{c_p + \frac{K}{Q}}{r}. \tag{4.4}$$

Let $\hat{q}_s$ denote the unique solution to this optimality condition and let $\hat{w}_s = r(1 - F(\hat{q}_s))$ denote the associated optimal price.

We next solve the manufacturer's problem under the 3D printer option and characterize the equilibrium of the Stackelberg game.

Proposition 4.2. *If* $c_p + \frac{K}{Q} \leq c_m$ *and* $\left(\min\{\hat{w}_s, w_m^A\} - c_p - \frac{K}{Q}\right) q_{nv}(\min\{\hat{w}_s, w_m^A\}) - S \geq \pi_M^b$*, then the manufacturer adopts 3D printing and does not use traditional manufacturing (i.e.,* $q_m = 0$*). The equilibrium wholesale price is* $\min\{\hat{w}_s, w_m^A\}$*, the order quantity is* $q_{nv}(\min\{\hat{w}_s, w_m^A\})$*, and the number of 3D printers utilized is* $q_{nv}(\min\{\hat{w}_s, w_m^A\})/Q$*. The manufacturer's profit is* $(\min\{\hat{w}_s, w_m^A\} - c_p - \frac{K}{Q}) q_{nv}(\min\{\hat{w}_s, w_m^A\}) - S$ *and the retailer's profit is* $\pi_R^b(r, \min\{\hat{w}_s, w_m^A\})$*. Otherwise, 3D printing is not utilized and the equilibrium is given in Corollary 4.1.*

Proof: See Appendix.

Example 4.1. *We explore the implications of Proposition 4.2 when* F *is the uniform distribution on* $[0, U]$*. Straightforward analysis shows that* $w_s = \frac{r + c_m}{2}$ *and* $\hat{w}_s = \frac{r + c_p + K/Q}{2}$*. For simplicity, let* $A = 0$*, which implies that* $w_m^A = r$ *and* $w_b = w_s$*. The conditions of the proposition, where 3D printing is used, can be written as* $c_p + \frac{K}{Q} \leq c_m$ *and*

$$\left(\hat{w}_s - c_p - \frac{K}{Q}\right) q_{nv}(\hat{w}_s) - S \geq (w_s - c_m) q_{nv}(w_s) \tag{4.5}$$

$$\Leftrightarrow \quad \left(\frac{r - c_p - K/Q}{2}\right) q_{nv}\left(\frac{r + c_p + K/Q}{2}\right) - \left(\frac{r - c_m}{2}\right) q_{nv}\left(\frac{r + c_m}{2}\right) \geq S \tag{4.6}$$

$$\Leftrightarrow \quad \frac{r}{4}\left(\left(1 - \frac{c_p + K/Q}{r}\right)^2 - \left(1 - \frac{c_m}{r}\right)^2\right) U \geq S. \tag{4.7}$$

Note that, since we require $c_p + \frac{K}{Q} \leq c_m$*, the upper bound on* S *is positive, and is increasing as both* c_p *and* K *decrease. Thus, as long as the fixed cost* S *of producing 3D printer plans is not too large, 3D printers will be purchased and utilized by the manufacturer. Also, as* $c_p + \frac{K}{Q}$ *decreases, we expect to see more manufacturer switching from traditional manufacturing to 3D printing.*

Finally, we compare the manufacturer's and retailer's equilibrium profits with those of a system with no 3D printing. The following corollary shows that, if 3D printing is economically feasible for the manufacturer to utilize, then it benefits *both* the manufacturer and retailer, with respect to the benchmark system in Section 4.1 where 3D printing is not available. Thus, the manufacturer's option of utilizing 3D printing can ameliorate the double marginalization effect.

Corollary 4.3. *If* $c_p + \frac{K}{Q} \leq c_m$, $\left(\min\{\hat{w}_s, w_m^A\} - c_p - \frac{K}{Q}\right) q_{nv}(\min\{\hat{w}_s, w_m^A\}) - S \geq \pi_M^b$ *and* $\min\{\hat{w}_s, w_m^A\} \leq w_b$, *then the manufacturer adopts 3D printing and both the manufacturer and retailer earn at least as much profit as the benchmark case of Corollary 4.1; if the second inequality is strict, the manufacturer earns strictly more profit; if the third inequality is strict, the retailer earns strictly more profit. Otherwise, both firms earn the same profits as described in Corollary 4.1.*

4.3 Retailer May Purchase 3D Printers

In this section we propose and analyze a novel scenario, where the retailer can potentially purchase 3D printers at the same cost K and unit printing cost c_p. We assume that the manufacturer sells both the physical product as well as 3D plans/schematics to the retailer, where in the latter case the retailer must print the product himself. If the manufacturer sells 3D schematics, she must incur a fixed cost of S to develop the schematics. In our game theoretic model, the manufacturer first offers the pricing for both fully manufactured products as well as the schematics for 3D printing: 1) the unit wholesale price of the manufactured products is w_m and 2) the unit 3D printing schematic prices w_p; the retailer paying for the 3D printing schematics compensates the manufacturer for the fixed cost of developing them. The retailer then decides 1) how much of the manufactured products to order q_m at wholesale price w_m, 2) how many 3D printers to purchase n, and 3) conditional on buying

at least one 3D printer, how many of the schematics q_p to purchase at w_p in order to 3D print them at unit cost c_p. Each 3D printer has the capacity to produce $Q > 0$ units over a fixed production horizon. Finally, since the manufacturer sets the pricing structure first, she is the leader, and the retailer is the follower in a Stackelberg game.

Remark 4.1. *In our proposed model, one potential concern for the manufacturer would be how to prevent the retailer from printing more than q_p products once the 3D schematics are provided. In order to tackle this issue, there are some software programs, such as PaperCut (www.papercut.com), that control the number of (paper) prints for each user, and the same technology could potentially be utilized in 3D printing. Moreover, many universities audit the printing activities of students and limit them. Thus, the technology for preventing the retailer from printing more than q_p is available.*

4.3.1 Retailer Analysis

The retailer maximizes his expected profit with respect to the decisions q_m, q_p, and n for a given (w_m, w_p) pricing structure. If he decides not to buy any 3D printers, q_p must be zero. The retailer's economics are as follows: as described above, (w_m, w_p) is the manufacturer's pricing structure, K is the fixed cost per 3D printer, c_p is the variable 3D printing cost, and r is the unit revenue. We assume that stochastic demand D is first fulfilled from the existing stock q_m (which arrives before the selling season), to avoid having left-over inventory, and remaining demand is fulfilled by the 3D printers, if any, up to their collective capacity nQ. In other words, our model is capturing and measuring the 3D printers' ability to increase the retailer's flexibility to satisfy demand. Therefore, we allow q_p to depend on the realized demand D; being closer to the consumer, it is reasonable to assume that this retailer decision can be made once demand uncertainty is resolved. Consequently, the amount of product that is 3D printed is represented by $q_p(D)$, which is limited by the total 3D printing capacity

nQ, as well as the remaining demand to be satisfied $\max\{D - q_m, 0\}$; this forms the main constraint in the retailer's profit-maximization problem:

$$\begin{aligned} \pi_R^{3D}(w_m, w_p) = \max_{q_m, q_p, n \geq 0} \quad & E[r \min\{q_m, D\}] - w_m q_m + E[(r - w_p - c_p) q_p(D)] - nK \\ \text{s.t.} \quad & q_p(D) \leq \min\{nQ, \max\{D - q_m, 0\}\}. \end{aligned} \tag{4.8}$$

The first two terms in the objective function are collectively the retailer's expected profit from manufactured products, and the third and fourth terms are the expected profit from 3D printed products. To avoid trivial solutions, we assume that $\max\{w_m, w_p + c_p\} < r$, so that purchasing either/both manufactured products and 3D schematics is economically feasible. The solution to the retailer's problem is presented in the following two propositions.

Proposition 4.3. *The optimal solution to Problem* (4.8)*, given the manufacturer's pricing structure* (w_m, w_p)*, is*

$$(q_m^*, n^*) = \begin{cases} \left(F^{-1}\left(1 - \frac{w_m}{r}\right), 0\right), & \frac{w_m}{r}(r - w_p - c_p) < \frac{K}{Q} \\ \left(F^{-1}\left(1 - \frac{w_m - \frac{K}{Q}}{w_p + c_p}\right), \frac{1}{Q}\left(F^{-1}\left(1 - \frac{K}{(r - w_p - c_p)Q}\right) - F^{-1}\left(1 - \frac{w_m - \frac{K}{Q}}{w_p + c_p}\right)\right)\right), & \begin{cases} w_m - w_p - c_p \leq \frac{K}{Q}, \\ \frac{K}{Q} \leq \frac{w_m}{r}(r - w_p - c_p), \end{cases} \\ \left(0, \frac{1}{Q} F^{-1}\left(1 - \frac{K}{(r - w_p - c_p)Q}\right)\right), & \frac{K}{Q} < w_m - w_p - c_p, \end{cases}$$

and $q_p^*(D) = \min\{n^* Q, \max\{0, D - q_m^*\}\}$.

Proof: See Appendix.

Proposition 4.4. *The maximized retailer profit is*

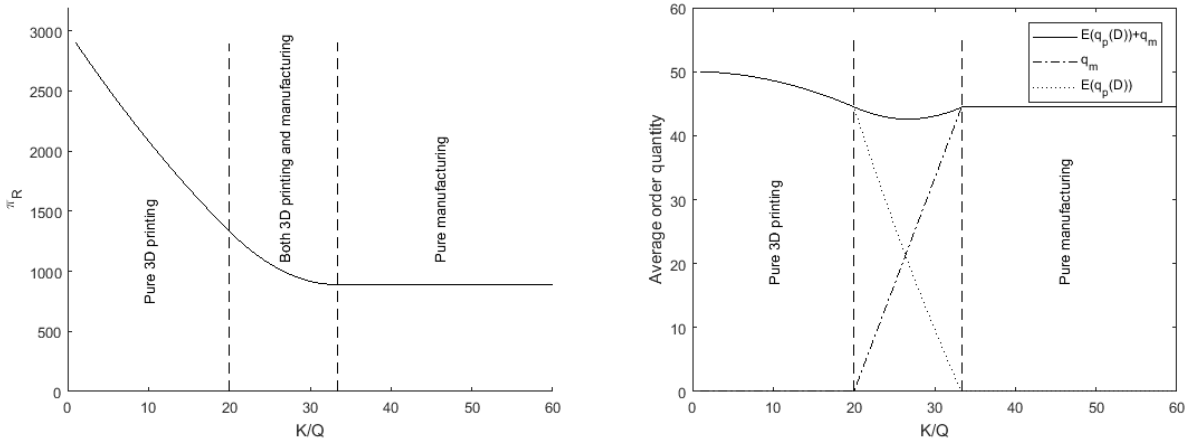
$$\pi_R^{3D}(w_m, w_p) = \begin{cases} \pi_R^b(r, w_m), & \frac{w_m}{r}(r - w_p - c_p) < \frac{K}{Q} \\ \pi_R^b\left(w_p + c_p, w_m - \frac{K}{Q}\right) + \pi_R^b\left(r - w_p - c_p, \frac{K}{Q}\right), & w_m - w_p - c_p \leq \frac{K}{Q} \leq \frac{w_m}{r}(r - w_p - c_p) \\ \pi_R^b\left(r - w_p - c_p, \frac{K}{Q}\right), & \frac{K}{Q} < w_m - w_p - c_p. \end{cases}$$

Proof: See Appendix.

Remark 4.2. *Note that our modeling of the retailer adoption of 3D printing is slightly different than that of the manufacturer adopting 3D printing. In the latter case, the printed amount of product is q_p, which does not depend on the demand D, whereas in the former, $q_p(D)$ is a function of D. The reason for this is that the manufacturer has a transportation lead time, where shipping typically takes place before demand is realized. However, in the former case, there is no transportation required for the products that are 3D printed at the retailer, and hence production can take place later to take advantage of realized demand, which motivates our modeling of $q_p(D)$.*

Propositions 4.3 and 4.4 have been presented in terms of the amortized fixed cost of a 3D printer K/Q. In the first case, where $\frac{w_m}{r}(r - w_p - c_p) < \frac{K}{Q}$, the 3D printers are too expensive and are not utilized. In the second case, where $w_m - w_p - c_p \leq \frac{K}{Q} \leq \frac{w_m}{r}(r - w_p - c_p)$, the 3D printers have a moderate cost, and they are utilized in parallel with traditional manufacturing, which is used less than in the first case (i.e., $F^{-1}\left(1 - (w_m - K/Q)/(w_p + c_p)\right) \leq F^{-1}\left(1 - w_m/r\right)$). In the third case, where $\frac{K}{Q} < w_m - w_p - c_p$, the 3D printers are cheap enough such that all production is 3D printed, and traditional manufacturing is not used. Since the cost K/Q is decreasing as we move from the first to the third case, the maximized profits are increasing in this order; we observe this visually in Figure 4.1a. In Figure 4.1b we observe the q_m and $E[q_p(D)]$ quantities, and notice that 1) in the pure 3D printing region,

$E[q_p(D)]$ is decreasing slowly, 2) in the hybrid region, q_m is increasing and $E[q_p(D)]$ is decreasing, and 3) in the pure manufacturing region, q_m is constant (since it doesn't depend on K/Q). Finally, note that the hybrid region found here is not possible in the case where the manufacturer adopts 3D printing, as shown in Section 4.2. For the rest of this subsection, we explore the situation where manufacturer is price taker. In other words, wholesale prices are determined exogenously. Note that $q_m^* + E(q_p^*(D))$ is the total demand that can be satisfied by the retailer.



(a) Retailer profit (b) Retailer order quantities

Figure 4.1: Retailer behavior. $r = 90, w_m = 50, w_p = 20, c_p = 10$.

Proposition 4.5. *For a uniform demand distribution:*

$$q_m^*+E(q_p^*(D)) = \begin{cases} 1 - \frac{w_m}{r}, & \frac{w_m}{r}(r - w_p - c_p) < \frac{K}{Q} \\ \frac{1}{2}\left[(\frac{w_m - K/Q}{w_p + c_p})^2 - (\frac{K/Q}{r - w_p - c_p})^2\right] + 1 - \frac{w_m - K/Q}{w_p + c_p}, & w_m - w_p - c_p \leq \frac{K}{Q} \leq \frac{w_m}{r}(r - w_p - c_p) \\ \frac{1}{2}\left[(\frac{w_m - K/Q}{w_p + c_p})^2 - (\frac{K/Q}{r - w_p - c_p})^2\right], & \frac{K}{Q} < w_m - w_p - c_p. \end{cases}$$

It is worth mentioning that in the second case, $\frac{\partial q_m^* + E(q_p^*(D))}{\partial w_m} = 0$. The interpretation is

that for larger values of w_m, the retailer orders less of the manufactured products. At the same time, he can 3D print the product more. These two effects cancel each other, and result in retailer being insensitive to w_m.

Next, notice that $E\left(D - [q_m^* + E(q_p^*(D))]\right)^+$ is the average mismatch between supply and demand, which is studied in the next corollary.

Corollary 4.4. *$q_m^* + E(q_p^*(D))$ is not monotone in Q.*

Transshipment literature (e.g., Dong and Rudi (2004), Proposition 1) claims more flexibility at retailer results in better balancing supply and demand. However, the intuition for the result of Corollary 4.4 is that with 3D printing (i.e., more flexibility), supply and demand are not necessarily better matched. In fact, there are regions where 3D printing increases mismatch between supply and demand. One example is shown in Figure 4.1b where in the hybrid region $q_m^* + E(q_p^*(D))$ is less than q_m^* in the pure manufacturing region. The reason is three fold. 1- Low 3D plan price compared to manufacturing variable cost. 2- Low 3D printing capacity. 3- High 3D printer fixed cost. In such a case, 3D printing is desirable for the retailer to use, however, due to its high fixed cost and low capacity, production is limited. Therefore, in this case, under mentioned conditions, while retailer's profit increases in presence of 3D printing, supply and demand matching decreases.

Corollary 4.5. *3D printing might not be adopted even if $w_p + c_p < w_m$.*

Emergency sourcing literature assumes that if the source with longer lead time is costlier than the source with shorter lead time, the former will not be used at all. One can observe that in Proposition 4.5, however, 3D printing (i.e., the source with zero lead time) might not even be adopted if $w_p + c_p < w_m$. This takes place due to the unique cost structure of this problem. That is, since the retailer opts for adopting 3D printing, its replenishment cost is incurred at the retailer. Therefore, if 3D printing variable cost is small, however, its

replenishment cost is large, it will not be adopted, and all the products will be ordered form the manufacturer.

4.3.2 Manufacturer Analysis

Since the manufacturer is the leader in the Stackelberg game, we assume she knows the retailer's best responses (i.e., q_m^*, $q_p^*(D)$, and n^*, as a function of w_m and w_p) when she makes her pricing decisions w_m and w_p. The manufacturer's costs are the unit manufacturing cost c_m for the fully complete units the retailer orders and, if the retailer chooses to buy 3D printing schematics, the fixed cost S to develop these schematics. We again capture exogenous economic factors via the retailer's opportunity cost $A \geq 0$, where the retailer will not participate unless he is assured his expected profit is at least A. The manufacturer's profit-maximization problem is

$$\begin{aligned} \pi_M^{3D(R)} = \max_{w_m, w_p \geq 0} \quad & (w_m - c_m) q_m^* + w_p E[q_p^*(D)] - S \cdot \mathbb{1}\{n^* > 0\} \\ \text{s.t.} \quad & (q_m^*, q_p^*(D), n^*) \text{ as defined in Proposition 4.3} \\ & \pi_R^{3D}(w_m, w_p) \geq A, \end{aligned} \tag{4.9}$$

where $\mathbb{1}\{\}$ is the indicator function. The first and second terms in the objective function represent the manufactured products and 3D schematics profits, respectively, and the third term is the fixed cost of developing 3D schematics (if $n^* > 0$). The first constraint is the retailer's best response, from Proposition 4.3, and the second constraint ensures the retailer's participation.

We were unable to analytically solve Problem (4.9) for a generic distribution F (even under the assumption that F has an increasing generalized failure rate). However, we were able to solve it for the case where demand is uniformly distributed on $[0, U]$, the retailers opportunity cost $A = 0$, and the cost to generate 3D plans $S = 0$; we present these results

next. Subsequently, we provide numerical experiments for the cases where demand is a truncated normal distribution, $A > 0$, $S > 0$, and observe results that are consistent with our analytical outcomes for the uniformly distributed demand case.

The three conditions of the retailer's best response, from Proposition 4.3, partition the (w_m, w_p) space into three regions. To simplify the exposition, we introduce notation for the three regions:

$$P_1 = \left\{(w_m, w_p) : \frac{K}{Q} > \frac{w_m}{r}(r - w_p - c_p)\right\}$$
$$P_2 = \left\{(w_m, w_p) : w_m - w_p - c_p \leq \frac{K}{Q} \leq \frac{w_m}{r}(r - w_p - c_p)\right\}$$
$$P_3 = \left\{(w_m, w_p) : \frac{K}{Q} < w_m - w_p - c_p\right\}.$$

Similarly, we define subproblems that are indexed by $i = 1, \ldots, 3$:

$$\begin{aligned} \pi_M^i = \max_{w_m, w_p \geq 0} \quad & (w_m - c_m)q_m^* + w_p E[q_p^*(D)] - S \cdot \mathbb{1}\{n^* > 0\} \\ \text{s.t.} \quad & (w_m, w_p) \in P_i \\ & (q_m^*, q_p^*(D), n^*) \text{ as defined in Proposition 4.3} \\ & \pi_R^{3D}(w_m, w_p) \geq 0, \end{aligned}$$

where $\pi_M^{3D(R)} = \max_{1 \leq i \leq 3} \pi_M^i$. We solve π_M^i, $i = 1, \ldots, 3$, in the following sequence of lemmas.

Lemma 4.1. *For uniformly distributed demand on* $[0, U]$ *and* $(A, S) = (0, 0)$, $(w_m^*, w_p^*) = \left(\frac{r + c_m}{2}, r - c_p\right)$ *and* $\pi_M^1 = \frac{1}{r}(\frac{r - c_m}{2})^2 U$. *Furthermore,* $(q_m^*, n^*, q_p^*) = \left(\left(\frac{r - c_m}{2r}\right) U, 0, 0\right)$ *and* $\pi_R^{3D}(w_m^*, w_p^*) = \frac{3}{2}\left(\frac{r - c_m}{2r}\right)^2 Ur$.

Proof: See Appendix.

We next present the results of the analysis of π_M^2; in order to effectively present them, we introduce some definitions that depend only on problem data, but not any decisions by

either firm:

$$\hat{y} \triangleq \frac{U}{(r+c_p)}\left(\frac{K}{2Q}+\frac{(r+c_p)}{2}-c_m+c_p+\sqrt{\left(\frac{K}{2Q}+\frac{(r+c_p)}{2}-c_m+c_p\right)^2+2(r+c_p)\left(c_m-c_p-\frac{K}{Q}\right)}\right)$$

and

$$\begin{aligned}
\Delta &\triangleq (r+c_m)^2-4(r+c_p)\frac{K}{Q}\\
\xi_1 &\triangleq \frac{(r+2c_p-c_m-\sqrt{\Delta})U}{2(r+c_p)} \text{ (for non-negative } \Delta)\\
\xi_2 &\triangleq \frac{(r+2c_p-c_m+\sqrt{\Delta})U}{2(r+c_p)} \text{ (for non-negative } \Delta)\\
\psi &\triangleq U-\frac{U}{6(r-c_p)}\left(\frac{K}{Q}+\left(\frac{K}{Q}\left(\sqrt{27(r-c_p)^2+\frac{K^2}{Q^2}}-\sqrt{27}(r-c_p)\right)^2\right)^{\frac{1}{3}}+\frac{\frac{K^2}{Q^2}}{\left(\frac{K}{Q}\left(\sqrt{27(r-c_p)^2+\frac{K^2}{Q^2}}-\sqrt{27}(r-c_p)\right)^2\right)^{\frac{1}{3}}}\right)\\
\zeta_1 &\triangleq \max\{0,\min\{\psi,\hat{y}\}\}\\
\zeta_2 &\triangleq \max\left\{\hat{y},\left(\frac{r-c_m}{2r}\right)U\right\}
\end{aligned}$$

$$\begin{aligned}
\hat{\zeta}_1 &\triangleq \max\{0,\min\{\psi,\xi_2\}\} \text{ (for non-negative } \Delta)\\
\hat{\zeta}_2 &\triangleq \max\left\{\xi_2,\left(\frac{r-c_m}{2r}\right)U\right\} \text{ (for non-negative } \Delta)\\
\phi &\triangleq \text{the unique root of } \frac{(\frac{K}{Q}-c_m+c_p)^2U^2\frac{K}{Q}}{2\left(\frac{KU}{Q}+(U-y)(r+c_p)\right)^2}+\frac{(r-c_p)(U-y)}{U}-\frac{K}{Q}-\frac{\frac{K}{Q}y(2U-y)}{2(U-y)^2}=0\\
&\quad \text{(which exists when } (K/Q-c_m+c_p)^2K/(2Q\left(K/Q+r+c_p\right)^2)+(r-c_p)\geq K/Q)\\
\chi_1 &\triangleq \max\{\xi_1,\min\{\phi,\xi_2\}\} \text{ (for non-negative } \Delta \text{ and real } \phi)\\
\chi_2 &\triangleq \begin{cases}\xi_1, & \left(\frac{r-c_m}{2r}\right)U\leq\xi_1\\ \xi_2, & \left(\frac{r-c_m}{2r}\right)U\geq\xi_1\\ \arg\max\limits_{y\in\{\xi_1,\xi_2\}}\Gamma(y,y). & \end{cases} \text{ (for non-negative } \Delta \text{ and real } \phi)
\end{aligned}$$

It is also convenient to make the following change of variables: $x = \left(1 - \frac{w_m - \frac{K}{Q}}{w_p + c_p}\right) U$, which is q_m in the second case of Proposition 4.3, and $y = \left(1 - \frac{K}{(r - w_p - c_p)Q}\right) U$, where $n = (y - x)/Q$ in the second case of Proposition 4.3. The prices can be recovered using the inverse functions $w_p = r - c_p - \frac{K}{Q(1-y/U)}$ and $w_m = r(1 - x/U) - \frac{K(1-x/U)}{Q(1-y/U)} + \frac{K}{Q}$.

Lemma 4.2. *For uniformly distributed demand on* $[0, U]$ *and* $(A, S) = (0, 0)$*:*

1. *If* $(r + c_m)^2 - 4(r + c_p)\frac{K}{Q} < 0$*, then* $y^* = \left(\frac{r - c_m}{2r}\right) U$*,* $x^* = y^*$*, and the manufacturer profit is* $\pi_M^2 = \frac{1}{r}\left(\frac{r - c_m}{2}\right)^2 U$*.*

2. *If* $(r + c_m)^2 - 4(r + c_p)\frac{K}{Q} \geq 0$*, then:*

   (a) *If* $\frac{K}{Q} - c_m + c_p \geq 0$*, then:*

   i. *If* $\frac{(\frac{K}{Q} - c_m + c_p)^2 \frac{K}{Q}}{2\left(\frac{K}{Q} + (r + c_p)\right)^2} + (r - c_p) < \frac{K}{Q}$*, then* $y^* = \xi_1$*,* $x^* = \left(\frac{K}{Q} - c_m + c_p\right) / \left(\frac{r + c_p}{U} - \frac{K}{Q(U - \xi_1)}\right)$*, and the manufacturer profit is* $\pi_M^2 = \frac{1}{2}\frac{\left(\frac{K}{Q} - c_m + c_p\right)^2}{\left(\frac{r + c_p}{U} - \frac{K}{Q(U - \xi_1)}\right)} + \left(r - c_p - \frac{KU}{Q(U - \xi_1)}\right)\left(\xi_1 - \frac{\xi_1^2}{2U}\right)$*.*

   ii. *If* $\frac{(\frac{K}{Q} - c_m + c_p)^2 \frac{K}{Q}}{2\left(\frac{K}{Q} + (r + c_p)\right)^2} + (r - c_p) \geq \frac{K}{Q}$*, then*
   $y^* = \begin{cases} \chi_1, & -\frac{1}{2}\frac{\left(\frac{K}{Q} - c_m + c_p\right)^2}{\left(\frac{K}{Q(U - \chi_1)} - \frac{(r + c_p)}{U}\right)} - \left(\frac{KU}{Q(U - \chi_1)} - r + c_p\right)\left(\chi_1 - \frac{\chi_1^2}{2U}\right) \geq \left(\frac{r}{U}(U - \chi_2) - c_m\right)\chi_2 \\ \chi_2, & \textit{otherwise.} \end{cases}$,
   $x^* = \begin{cases} -(\frac{K}{Q} - c_m + c_p)/(\frac{K}{Q(U - y^*)} - \frac{r}{U} - \frac{c_p}{U}), & \xi_1 < y^* < \xi_2 \\ y^*, & y^* \leq \xi_1 \textit{ or } y^* \geq \xi_2, \end{cases}$ *and the manufacturer profit is*
   $\pi_M^2 = -\frac{1}{2}\left(\left(\frac{K}{Q} - c_m + c_p\right)^2\right) / \left(\left(\frac{K}{Q(U - y^*)} - \frac{(r + c_p)}{U}\right)\right) + \left(r - c_p - \frac{KU}{Q(U - y^*)}\right)\left(y^* - \frac{(y^*)^2}{2U}\right)$*.*

   (b) *If* $\frac{K}{Q} - c_m + c_p < 0$*, then:*

   i. *If* $\xi_2 \leq \hat{y}$*, then* $y^* = \begin{cases} \zeta_1, & \left(r - c_p - \frac{KU}{Q(U - \zeta_1)}\right)\left(\zeta_1 - \frac{\zeta_1^2}{2U}\right) \geq \left(\frac{r}{U}(U - \zeta_2) - c_m\right)\zeta_2 \\ \zeta_2, & \textit{otherwise.} \end{cases}$, $x^* = \begin{cases} 0, & y^* \in [0, \hat{y}] \\ y^*, & y^* \in (\hat{y}, U]. \end{cases}$ *and the manufacturer profit is* $\pi_M^2 = \begin{cases} \left(r - c_p - \frac{KU}{Q(U - y^*)}\right)\left(y^* - \frac{(y^*)^2}{2U}\right), & y^* \in [0, \hat{y}] \\ \left(\frac{r}{U}(U - y^*) - c_m\right) y^*, & y^* \in (\hat{y}, U] \end{cases}$.

ii. If $\xi_2 > \hat{y}$, *then* $y^* = \begin{cases} \hat{\zeta}_1, & \left(r - c_p - \frac{KU}{Q(U-\hat{\zeta}_1)}\right)\left(\hat{\zeta}_1 - \frac{\hat{\zeta}_1^2}{2U}\right) \geq \left(\frac{r}{U}(U-\hat{\zeta}_2) - c_m\right)\hat{\zeta}_2 \\ \hat{\zeta}_2, & \textit{otherwise.} \end{cases}$,

$x^* = \begin{cases} 0, & y^* \in [0, \xi_2] \\ y^*, & y^* \in (\xi_2, U]. \end{cases}$, *and the manufacturer profit is*

$\pi_M^2 = \begin{cases} \left(r - c_p - \frac{KU}{Q(U-y^*)}\right)\left(y^* - \frac{(y^*)^2}{2U}\right), & y^* \in [0, \xi_2] \\ \left(\frac{r}{U}(U-y^*) - c_m\right)y^*, & y^* \in (\xi_2, U] \end{cases}$.

In all cases, the retailer profit is

$$\pi_R^{3D}(w_m^*, w_p^*) = \frac{K}{2Q(U-y^*)}((y^*)^2 - (x^*)^2) + \frac{r(x^*)^2}{2U}.$$

Proof: See Appendix.

Lemma 4.3. *For uniformly distributed demand on* $[0, U]$ *and* $(A, S) = (0, 0)$, $y^* = \max\{0, \psi\}$, $x^* = 0$, *the manufacturer profit is* $\pi_M^3 = \left(r - c_p - \frac{KU}{Q(U-y^*)}\right)\left(y^* - \frac{(y^*)^2}{2U}\right)$ *and the retailer profit is* $\pi_R^{3D}(w_m^*, w_p^*) = \frac{(y^*)^2 K}{2Q(U-y^*)}$.

Proof: See Appendix.

The following proposition assembles and simplifies the results from Lemmas 4.1 - 4.3.

Proposition 4.6. *For uniformly distributed demand on* $[0, U]$ *and* $(A, S) = (0, 0)$*:*

1. *If* $(r + c_m)^2 - 4(r + c_p)\frac{K}{Q} < 0$, *then* $\max\{\pi_M^1, \pi_M^3\}$ *determines the equilibrium.*

2. *If* $(r + c_m)^2 - 4(r + c_p)\frac{K}{Q} \geq 0$, *then:*

   (a) *If* $\frac{K}{Q} - c_m + c_p \geq 0$, *then* $\max\{\pi_M^1, \pi_M^2, \pi_M^3\}$ *determines the equilibrium.*

   (b) *If* $\frac{K}{Q} - c_m + c_p < 0$, *then* $\max\{\pi_M^1, \pi_M^3\}$ *determines the equilibrium.*

Proof: See Appendix.

Proposition 4.6 has been proved, but is incomplete; unfortunately, we found it intractable to make more analytical progress. We next present our conjectured solution, which is subsequently supported by numerical experiments.

Conjecture 4.1. *For uniformly distributed demand on* $[0, U]$ *and* $(A, S) = (0, 0)$*:*

1. *If* $(r + c_m)^2 - 4(r + c_p)\frac{K}{Q} < 0$*, then the equilibrium is manufacturing.*

2. *If* $(r + c_m)^2 - 4(r + c_p)\frac{K}{Q} \geq 0$*, then:*

   (a) *If* $\frac{K}{Q} - c_m + c_p \geq 0$*, then:*

      i. *If* c_m *is not too large and* $c_p + K/Q$ *is not too small, then the equilibrium is manufacturing.*

      ii. *Otherwise, the equilibrium is a hybrid utilizing both manufacturing and 3D printing.*

   (b) *If* $\frac{K}{Q} - c_m + c_p < 0$*, then the equilibrium is 3D printing.*

In Figure 4.2 we numerically evaluate the equilibrium outcomes (pure 3D printing, pure manufacturing, and hybrid) for $(c_p, c_m) \in [0, r]^2$ for $K/Q \in \{20, 40\}$ and $r = 90$. We superimpose the conditions of Proposition 4.6, in order to compare the induced partition of (c_p, c_m) space with those of the actual equilibrium. Increasing K/Q further continues the dynamic apparent in Figure 4.2 (i.e., shrinking regions corresponding to cases 2a and 2b, and a growing region corresponding to case 1).

We observe that cases 1 and 2b of Proposition 4.6 exactly correspond to the actual equilibrium outcomes: in case 1, the equilibrium is pure manufacturing, and in case 2b, the equilibrium is pure 3D printing. However, in case 2a, there are two possible equilibria not precisely captured by our analysis: 1) pure manufacturing or 2) a hybrid of manufacturing

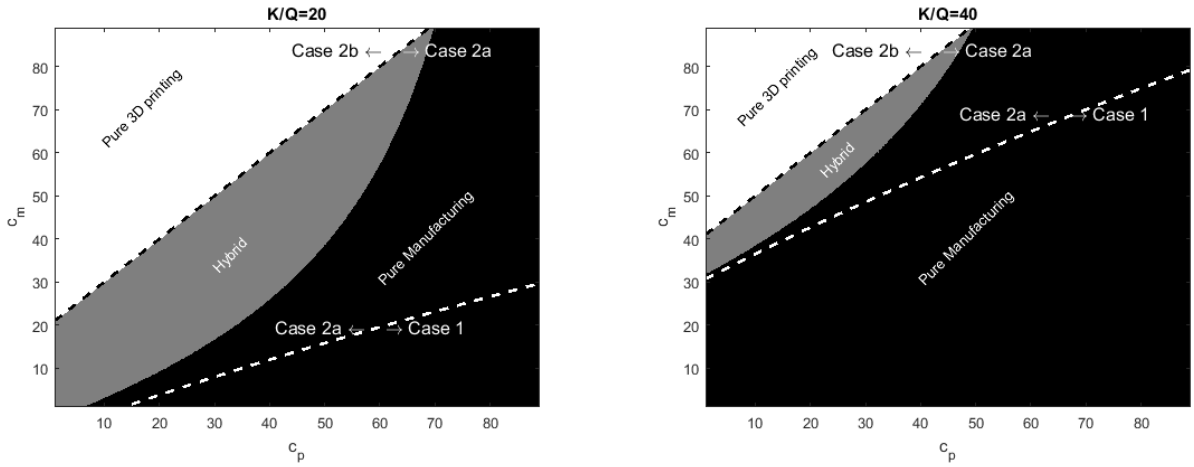


Figure 4.2: Equilibrium outcomes compared with cases of Proposition 4.6.

and 3D printing. We attempt to capture these equilibria in the sub-cases i–ii of case 2a in Conjecture 4.1, and we next provide a discussion of them.

The condition of case 1 can be manipulated to obtain a clear interpretation:

$$(r+c_m)^2 - 4(r+c_p)\frac{K}{Q} < 0 \Rightarrow \frac{r+c_m}{2} < \sqrt{(r+c_p)\frac{K}{Q}} \Rightarrow \frac{r+c_m}{2} < \frac{r+c_p+\frac{K}{Q}}{2},$$

where the second implication is due to the inequality of arithmetic and geometric means, and the last inequality can be simplified to $c_m < c_p + \frac{K}{Q}$. In other words, the condition of case 1 implies that the unit cost of traditional manufacturing is strictly less than the unit cost of 3D printing, which leads to pure manufacturing in equilibrium. In case 2a, the inequality $c_m \leq c_p + \frac{K}{Q}$ holds, but due to the restriction $(r+c_m)^2 - 4(r+c_p)\frac{K}{Q} \geq 0$, traditional manufacturing can not be too much cheaper than 3D printing, leading to a hybrid equilibrium in a portion of the 2a region; the use of 3D printing, despite its higher cost, is due to the additional flexibility it provides in reacting to demand after it is realized. Case 2b considers the region where 3D printing has a strictly smaller unit printing cost

than traditional manufacturing, $c_p + \frac{K}{Q} < c_m$, which results in the equilibrium being pure 3D printing. Finally, note that the additional cases i–ii for case 2a in Conjecture 4.1 are consistent with these interpretations.

In Figure 4.3, we provide the percentage improvement in the manufacturer's profit, with respect to the benchmark profit: $(\pi_M^{3D(R)} - \pi_M^b)/\pi_M^b$. The setup is similar to above, except we consider $(c_p, c_m) \in [20, 70]^2$, where $r = 90$, in order to eliminate extreme cases of very cheap or very expensive unit costs (which can result in 3D printing improving benchmark profits by an unrealistic multiple of 1000). From the graphs, it is evident that as c_m increases and c_p (or K/Q) decreases, the benefit of 3D printing improves, an intuitive finding. What is perhaps less intuitive is the growth rate of improvement: observing the scale of contour plot, we see that the benefit of 3D printing increases very fast as c_m decreases or c_p increases, approaching an improvement of approximately 900% when $(r, c_m, c_p, K/Q) = (90, 70, 20, 20)$. While this exact point might not be realistic, our results suggest that substantial increases in manufacturer profit, due to the adoption of 3D printing, are possible. In Figure 4.4, we

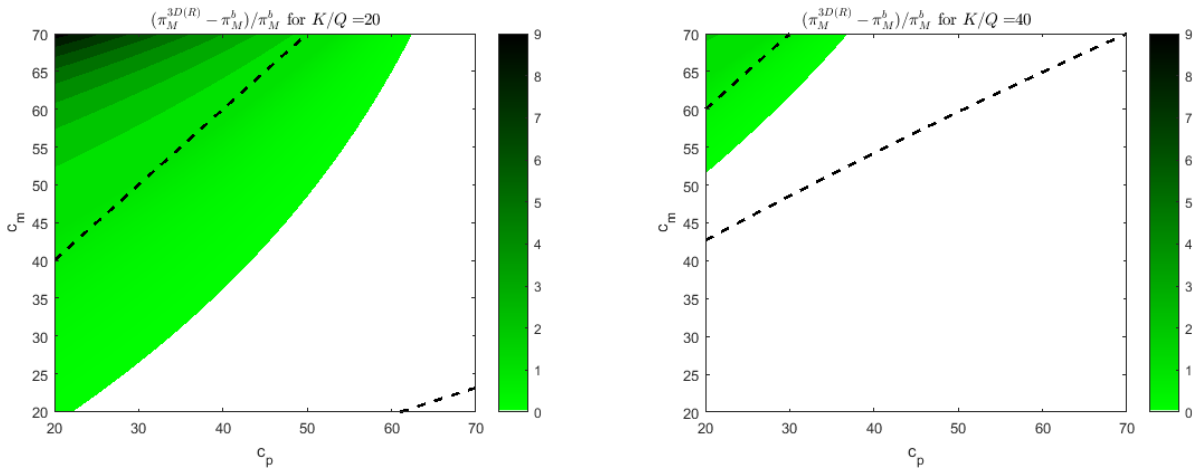


Figure 4.3: Ratios of manufacturer profit under 3D printing to benchmark manufacturer profit.

provide the percentage improvement in the retailer's profit, with respect to the benchmark profit: $(\pi_R^{3D} - \pi_R^b)/\pi_R^b$. We first observe that "win-win" situations are possible: when $c_p + K/Q$ is small and c_m is large (the top left corner of the left plot), the retailer's profit under 3D printing is strictly greater than its benchmark profit, just as in the manufacturer's case; however, the retailer's lift is not nearly as great as the manufacturer's. Unfortunately, the retailer can also lose a substantial amount of profit, whenever the equilibrium is a hybrid solution, as well as when the equilibrium is pure 3D printing with the additional condition that c_m is not too large with respect to $c_p + K/Q$. Therefore, the retailer benefits from 3D printing whenever the production economics are very much in its favor (i.e., when c_m is large with respect to $c_p + K/Q$). Of course, these negative outcomes can be prevented by increasing the retailer's reservation profit from $A = 0$ to a larger value; we shortly provide additional numerical results that study how the equilibrium changes as A is increased.

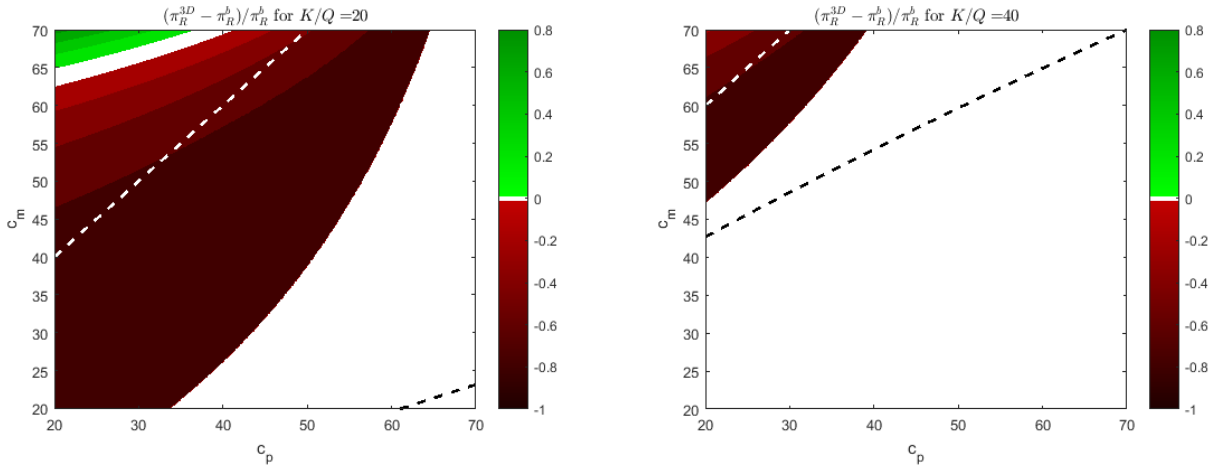


Figure 4.4: Ratios of retailer profit under 3D printing to benchmark retailer profit.

4.3.2.1 Relaxing the Assumptions of $S = 0$ and $A = 0$.

Figure 4.5 presents the optimal firm profits under 3D printing, as well as the optimal benchmark profits, as a function of S, the cost to develop 3D plans from traditional manufacturing plans, for $(A, K, Q, c_m, c_p) = (0, 150, 15, 40, 40)$. We observe that there exists a threshold value of S, after which both firm profits are independent of S since 3D printing is not the equilibrium (it is too expensive), and before which the manufacturer's profit is decreasing in S and the retailer's profit does not depend on S; therefore, when 3D printing forms part of the equilibrium, the cost to develop 3D plans is incurred solely by the manufacturer, and no costs are passed on to the retailer. However, note that, in this experiment, the benchmark manufacturer profit is 250, yet the manufacturer's profit under 3D printing is larger than this benchmark for S up to 300; note that this occurs even though the unit cost of manufacturing, $c_m = 40$, is strictly less than that of 3D printing, $c_p + K/Q = 50$. In addition, in passing through the threshold value of S, the manufacturer's profit is continuous, but the retailer experiences a substantial drop; near this transition point, the retailer may consider a lump sum payment to subsidize the manufacturer's cost S, in order to avoid the drop in profit.

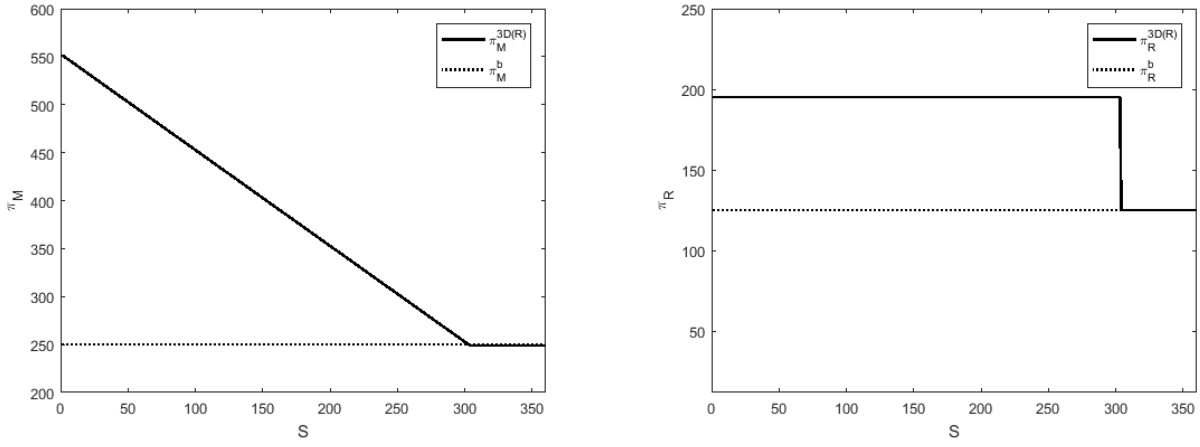


Figure 4.5: Optimal firm profits as a function of S.

Figure 4.6 presents the optimal firm profits under 3D printing, as well as the optimal benchmark profits, for different economic conditions, as represented by the retailer's opportunity cost A, for $(S, K, Q, c_m, c_p) = (0, 150, 15, 40, 40)$. We observe that there exists a threshold value of A, before which both firm profits are independent of A (the retailer is earning more than A), and after which the manufacturer's profit is decreasing in A and the retailer's profit is increasing in A; this threshold is slightly higher in the benchmark case. When A is large enough, the retailer is ambivalent between 3D printing and traditional manufacturing, whereas the manufacturer always prefers 3D printing.

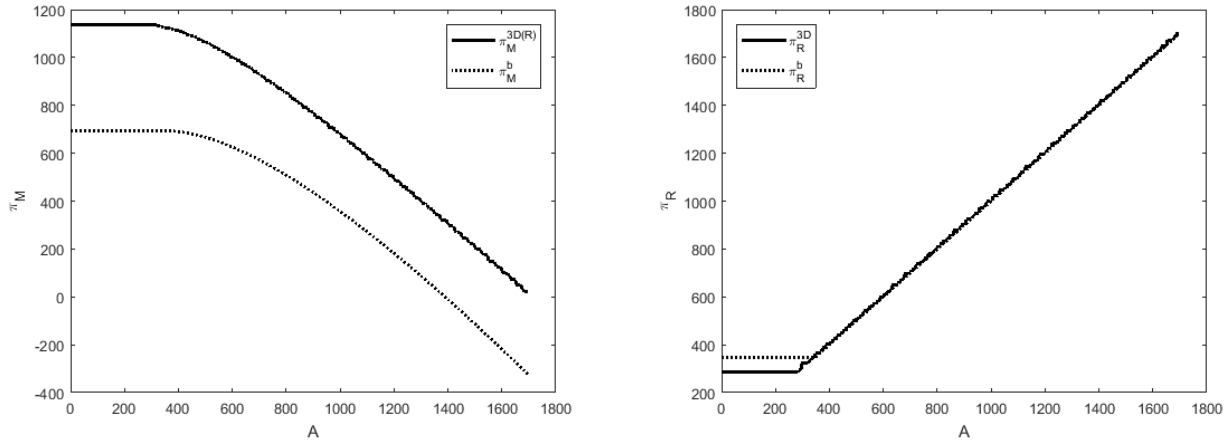


Figure 4.6: Optimal firm profits as a function of A.

4.3.2.2 Relaxing the Assumption of Uniform Demand.

In this section we consider the case where demand is normally distributed (truncated at zero), rather than uniformly distributed, and we (numerically) show that the equilibrium is qualitatively similar. Since we were unable to make analytical progress for the manufacturer's behavior for any distribution other than the uniform, we resorted to purely numerical evaluation of the equilibrium. We found the time requirements prohibitive to generate contour plots

for the normal distribution, as in Figures 4.2–4.4; we estimate 500 hours per contour plot for comparable resolution; thus, an additional contribution of our previous analytical results, for the uniform distribution, is the ability to efficiently generate Figures 4.2–4.4. Consequently, for the normal distribution of demand, we provide "slices" of the contour plots, which still allow comparisons with the uniform distribution results. In particular, we provide evidence that our analytical results are not fully dependent on the uniform distribution, and that the insights we generate are applicable to other distributions.

The experimental design is similar to that above, except that the demand is normally distributed, truncated at zero, with the same mean of $\mu = 50$, and the standard deviation is $\sigma = 15$. In Figure 4.7 we identify the equilibrium as a function of c_p for $c_m \in \{35, 60, 67.5\}$ and $K/Q = 20$; these values of c_m were selected in order to compare with qualitatively different behaviors in the uniform distribution case. Comparing with horizontal slices of Figure 4.2, at the same values of c_m considered here, we see that the equilibrium outcomes are effectively identical.

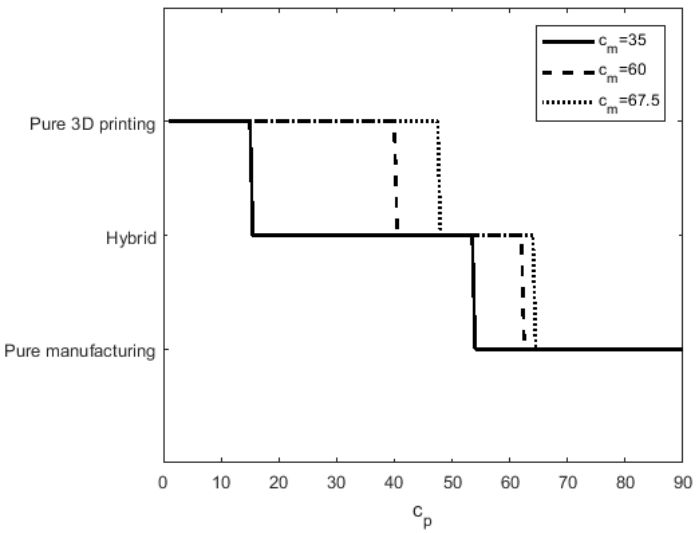


Figure 4.7: Equilibrium outcomes for a normal distribution of demand.

In Figure 4.8, we plot, on the left, $(\pi_M^{3D(R)} - \pi_M^b)/\pi_M^b$, and, on the right, $(\pi_R^{3D} - \pi_R^b)/\pi_R^b$, for the normal distribution of demand, which allow comparisons with slices of the left plots in Figures 4.3 and 4.4 (for $K/Q = 20$); the experimental setup is identical to that of Figure 4.7. For the manufacturer, we see identical behaviors in the uniform and normal cases: the ratio decreases in c_p to zero, where the equilibrium is manufacturing; furthermore, we observe that, in the left plots of both Figures 4.3 and 4.8, the slope is steeper for larger values of c_m. Similar behaviors are observed for the retailer when comparing the left plot of Figure 4.4 (where $K/Q = 20$) with the right plot of Figure 4.8. Consequently, we conjecture that most of the insights generated for the uniform distribution are qualitatively applicable to other distributions of demand.

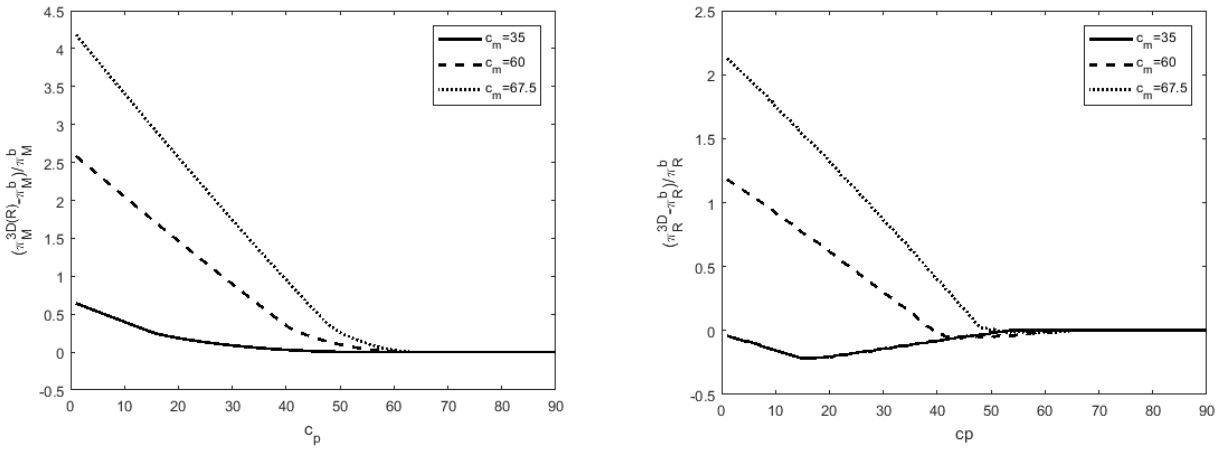


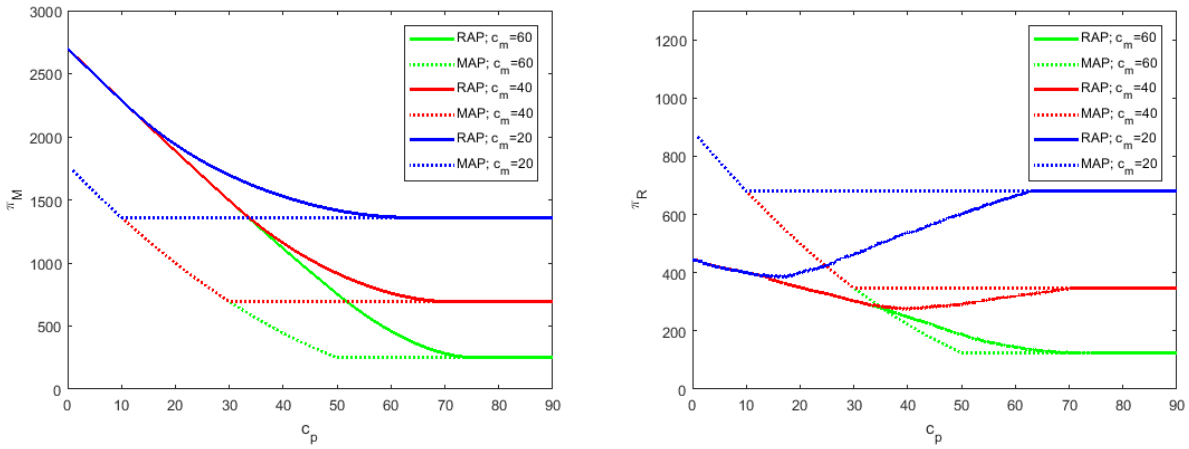
Figure 4.8: Ratios of firm profits under 3D printing to benchmark firm profits.

In conclusion, manufacturer, being the leader, always benefits from the presence of 3D printing option. In the worst case, if 3D printing is not beneficial for her, she sets w_m and w_p such that 3D printing is not adopted by the retailer. Otherwise, she sets w_m and w_p such that the retailer adopts 3D printing and capture more demand. The manufacturer, then, captures the extra profit from him in the equilibrium. The retailer, however, could benefit

or even *suffer* from the presence of 3D printing.

4.4 Manufacturer versus Retailer Adoption of 3D Printing

In this section we compare firm and system performance of the manufacturer adopting 3D printing (MAP) scenario, as analyzed in Section 4.2, with the retailer adopting 3D printing (RAP) scenario, as analyzed in Section 4.3. We utilize the same numerical setup as in the previous section for the uniform distribution; we also mention that the following results are qualitatively similar to results obtained under many different parameterizations and distributions (e.g., the truncated normal). Our study suggests that the manufacturer and the system always earn more profit under the RAP scenario; the retailer, in constrast, earns more profit under the MAP scenario, for most problem parameters. However, there are some problem parameters that lead to both firms preferring RAP.



(a) Manufacturer profit

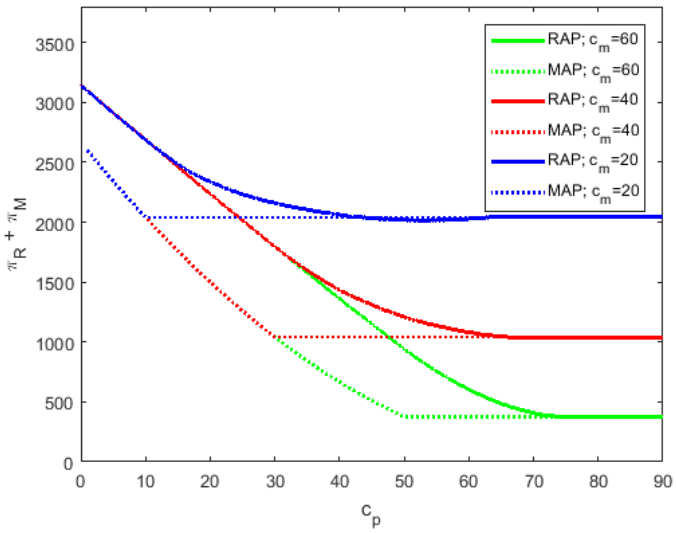
(b) Retailer profit

Figure 4.9: Optimal firms profits as a function of c_p for $S = 0$ and $c_m \in \{20, 40, 60\}$

Figures 4.9a and 4.9b depict the manufacturer and retailer profits as a function of c_p, for different values of c_m, when $(S, K, Q, A) = (0, 150, 15, 0)$. The solid and dashed line,

on each graph, represent each firm's profit under RAP and MAP, respectively. It is worth noting that when the curves become horizontal lines, the equilibrium does not adopt 3D printing. According to Figure 4.9a, for all values of c_p and c_m considered, RAP always results in more manufacturer profit than MAP. Figure 4.9b, on the other hand, shows that the retailer earns more profit under RAP only if c_m is large and c_p is moderate (i.e., within a range); otherwise, the retailer prefers MAP. Our study further suggests that the manufacturer conclusions remain unchanged (RAP dominates MAP) as S or $\frac{K}{Q}$ are varied; however, the retailer can prefer MAP for all (c_m, c_p) values if S or $\frac{K}{Q}$ are large enough. Finally, Figure 4.10 confirms that the system behavior mirrors that of the manufacturer.

Figure 4.10: System profit as a function of c_p for $S = 0$ and $c_m \in \{20, 40, 60\}$



4.5 Centralized System

In this section we propose and analyze the centralized system with presence of 3D printing option. If 3D printing is adopted, a fixed cost of S to develop the schematics must be incured by the system. The system decides 1) how many of the products to manufacture q_m at a

variable cost of c_m, 2) how many 3D printers to purchase n, and 3) conditional on buying at least one 3D printer, how many of the products q_p to 3D print at unit cost c_p. Each 3D printer has the capacity to produce $Q > 0$ units over a fixed production horizon. Finally, system's profit-maximization problem is

$$\begin{aligned} \pi_R^{3D}(w_m, w_p) = \max_{q_m, q_p, n \geq 0} \quad & E[r \min\{q_m, D\}] - c_m q_m + E[(r - c_p) q_p(D)] - nK - S \cdot \mathbb{1}\{n^* > 0\} \\ \text{s.t.} \quad & q_p(D) \leq \min\{nQ, \max\{D - q_m, 0\}\}. \end{aligned} \tag{4.10}$$

The first two terms in the objective function are collectively the system's expected profit from manufactured products, and the third and fourth terms are the expected profit from 3D printed products. The last term ensures if 3D printing is adopted, the fixed cost of developing 3D plans is incurred. The solution to the system's problem is presented in the following two propositions.

Proposition 4.7. *The optimal solution to Problem* (4.10) *is*

$$(q_m^*, n^*) = \begin{cases} \left(0, \frac{1}{Q} F^{-1}\left(1 - \frac{K}{(r-c_p)Q}\right)\right), & \frac{K}{Q} < c_m - c_p \wedge \pi_R^b\left(r - c_p, \frac{K}{Q}\right) \geq \pi_R^b(r, c_m) + S, \\ \left(F^{-1}\left(1 - \frac{c_m - \frac{K}{Q}}{c_p}\right), \frac{\Gamma}{Q}\right), & c_m - c_p \leq \frac{K}{Q} \leq \frac{c_m}{r}(r - c_p) \wedge \pi_R^b\left(c_p, c_m - \frac{K}{Q}\right) + \pi_R^b\left(r - c_p, \frac{K}{Q}\right) \geq \pi_R^b(r, c_m) + S, \\ \left(F^{-1}\left(1 - \frac{c_m}{r}\right), 0\right), & \textit{otherwise.} \end{cases}$$

where $\Gamma = \left(F^{-1}\left(1 - \frac{K/Q}{(r-c_p)}\right) - F^{-1}\left(1 - \frac{c_m - \frac{K}{Q}}{c_p}\right)\right)$ *and* $q_p^*(D) = \min\{n^* Q, \max\{0, D - q_m^*\}\}$.

Proposition 4.8. *The maximized System profit is*

$$\pi_R^{3D} = \begin{cases} \pi_R^b\left(r - c_p, \frac{K}{Q}\right) - S, & \frac{K}{Q} < c_m - c_p \wedge \pi_R^b\left(r - c_p, \frac{K}{Q}\right) \geq \pi_R^b(r, c_m) + S, \\ \pi_R^b\left(c_p, c_m - \frac{K}{Q}\right) + \pi_R^b\left(r - c_p, \frac{K}{Q}\right) - S, & \begin{cases} c_m - c_p \leq \frac{K}{Q} \leq \frac{c_m}{r}(r - c_p), \\ \pi_R^b\left(c_p, c_m - \frac{K}{Q}\right) + \pi_R^b\left(r - c_p, \frac{K}{Q}\right) \geq \pi_R^b(r, c_m) + S, \end{cases} \\ \pi_R^b(r, c_m), & \textit{otherwise.} \end{cases}$$

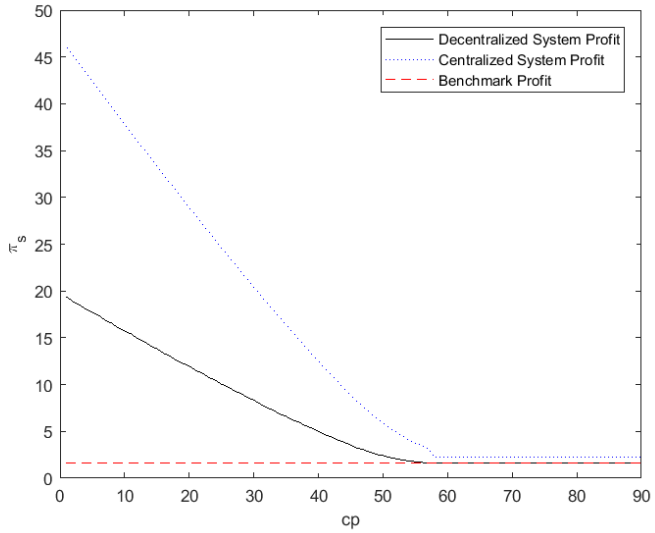
The three regions in Propositions 4.7 are similar to the ones in Proposition 4.3. Our numerical study shows that the upper bound on c_p and $\frac{K}{Q}$ before which 3D printing is adopted by the retailer under centralized system is greater compared to the decentralized system. Moreover, the lower bound on c_m, after which 3D printing is qadopted by the retailer under centralized system is less than decentralized system.

Figure 4.11 Shows that under presence of 3D printing, both centralized and decentralized systems are at least better off compared to the bench mark. However, centralized system's profit increase is greater than decentralized system's. In other words, while 3D printing option increases both system's profit, it worsens double marginalization effect.

4.6 Conclusion

In this chapter, the economic and competitive situations where either a retailer or a manufacturer, in a supply chain, would adopt 3D printing is characterized. The resulting profits are reported, for both firms, with respect to a benchmark supply chain without 3D printing. In case when the manufacturer opts for adopting 3D printing, both firms (i.e., the manufacturer and the retailer) are at least better off compared to the benchmark. Therefore, double marginalization effect is ameliorated. In case when the retailer opts for adopting 3D printing, the whole supply chain is more responsive to demand, because the retailer is closer to the

Figure 4.11: System profit comparison as a function of c_p for $(c_m, \frac{K}{Q}, S) = (70, 25, 0)$



final customers. That is, he can 3D print the products after the uncertainty of demand is realized. This problem is partially solved, and via a set of numerical studies, it is shown that the manufacturer is always at least better off. The retailer, however, might be better off or worst off depending on the problem parameters.

Chapter 5

CONCLUSIONS AND EXTENSIONS

One major question, which has long been studied in operations management, is how to efficiently match supply and demand. This is a crucial question because it has significant impact on cost management of a firm or a supply chain. As new technologies emerge, firms should update their matching supply-demand policy. In my dissertation, I focused on two emerging technologies, namely, cloud computing and 3D printing. In Chapter 3, with a specific focus on the former, markets, in which one unit of supply increases the capacity of two attributes simultaneously, were studied. The distinctive feature of this problem is that while demands for attributes are time-dependent, their ratio is disproportionate to supply ratio. In Chapter 4, simple supply chains, which consist of a retailer and a manufacturer, serving stochastic customers were studied. In one scenario the manufacturer is given the option to opt for 3D printing, and in another scenario, the retailer is given this opportunity. The latter can potentially decrease supply and demand mismatch in a supply chain.

5.1 Cloud Computing

In recent years, a huge surge in demand for cloud services has posed a challenge for the cloud infrastructure providers of how to expand capacities of computation resources efficiently to meet the surging demand. Specifically, the key determinants of configurations of physical and virtual machines, such as CPU and RAM, are often added in pre-configured supply packages (e.g., server clusters), instead of individual capacity scaling. Although one unit of a supply package bundles up a fixed ratio of the capacity attributes, the demand growth rates of these

attributes are time-varying and disproportionate. Hence, the optimal determination of the capacity-expansion policy, using the supply packages with bundled capacity attributes, has become a critical challenge.

First, we analyzed a family of capacity-expansion policies, for which the planning horizon is divided into multiple CECs. In each cycle, capacities of the two attributes are added through acquisitions of the two pre-configured cluster-types, either sequentially or simultaneously. Both attributes' capacities should be leveled at the beginning and at the end of the cycle. For every cycle, we derived the optimal timing and quantities of the expansions. Further, to determine the optimal lengths of the CECs, we devised two procedures: the *DP-algorithm* and the *FL-heuristic*. The FL-heuristic is to minimize the total cost rate of each CEC. The heuristic is easy to communicate and implement, and it can be used as an excellent starting point for the tedious DP-algorithm.

Second, cloud infrastructure providers desire to employ the minimum variety of the cluster-types for the sake of server fungibility. Thus, we examined the following question: When is the employment of only two cluster-types as good as the employment of a number of cluster-types?

To answer this question, we provided a lower bound for the total cost, including the fixed expansion and the holding costs. The lower bound results from an ideal policy, for which the provider can re-customize the configurations of supply packages at any expansion point, or, in other words, the provider can employ an infinite number of pre-configured cluster-types. We showed that, when the savings on the fixed expansion costs due to the joint replenishments are significant, the provider would adopt the simultaneous-replenishment policy, and the total cost would be close enough to that of the ideal policy. The implication is that, in such a case, employing two cluster-types must be good enough. In contrast, when the savings on the fixed expansion costs are insignificant, the provider may want to adopt the sequential-

replenishment policy. In such a case, the maximum benefit of employing more than two cluster-types cannot exceed the gap between the total cost of the sequential-replenishment policy with two cluster-types and that of the ideal policy. Then, we find critical conditions under which such a gap is relatively small. The important implication is that, under those conditions, the benefit of employing additional cluster-types is relatively small, and, thus, the employment of only two cluster-types can be good enough.

Third, we investigated the optimal selection of two cluster-types from many possible configurations. In addition to the *exhaustive search* approach to the cluster-selection problem, we also proposed a *cluster-selection heuristic* for the sequential-replenishment policy. The cluster-selection heuristic is as follows: Choose the configuration of the leading cluster that will lead to the minimal excess capacity when it is employed consecutively, and then choose the configuration of the following cluster that will lead to the fastest depletion of the excess capacity when it is employed in the end. Although the cluster-selection heuristic suggests the desirable ratios of capacity attributes for the two cluster-types, there could be multiple cluster-types that satisfy each ratio. In that case, among all cluster-types that satisfy the same ratio, one should choose the cluster-type with the greatest (smallest) number of attributes, if the unit purchase cost exhibits the economies (diseconomies) of scale. We showed that the performance of the cluster-selection heuristic is satisfactory.

Finally, we acknowledge that, as a preliminary study of the cloud computing capacity-expansion problems from the operations management area, this research has several limitations. Addressing these limitations can lead to valuable future research. On the demand side, researchers can consider stochastic demand growth rates and examine the optimal expansion policy. That said, researchers need to choose the assumptions about the consequences of capacity shortages in such a stochastic model, which makes the analysis of such a model complicated. Moreover, researcher also may want to consider a cloud provider's pricing de-

cisions in response to other providers' competitive strategies and, as a result, the impact on the demand forecasts and the provider's capacity-expansion decisions. On the supply side, although our study has provided useful conditions under which the employment of only two cluster-types is as efficient as the employment of many cluster-types, researchers can analyze capacity-expansion policies, using an arbitrary number of cluster-types. Such an analysis supplements our study, when those conditions that justify the employment of only two cluster-types are not satisfied.

5.2 Additive Manufacturing

In Chapter 4, we studied the impacts 3D printing, or additive manufacturing, can have on the performance of a simple supply chain consisting of a manufacturer and retailer. Due to the unique characteristics of 3D printing, either the manufacturer or retailer can adopt it. In this chapter we have characterized the economic and competitive situations where either firm would adopt 3D printing, and we report on the resulting profits, for both firms, with respect to a benchmark supply chain without 3D printing.

In the case when the manufacturer opts for adopting 3D printing technology, we find the equilibrium for a generic demand distribution. We show that both firms (i.e., the manufacturer and the retailer) are at least better off compared to a benchmark case with no 3D printing option available. Thus, this scenario ameliorates double marginalization effect. Furthermore, if all adopting-3D printing conditions are satisfied, all the products will be 3D printed; otherwise, they sill be traditionally manufactured.

In the case when the retailer opts for adopting 3D printing technology, for a generic demand distribution, we find retailer's best response to manufacturer's wholesale price strategy. Under a few simplification assumptions, we find the equilibrium for a uniform demand distribution, and through a set of numerical studies, we show that the manufacturer is always

at least better off. The retailer, however, might be better off or worst off depending on the problem parameter. Furthermore, using a combination of the traditionally manufactured product and 3D printing is possible in this case. We, finally, confirm the robustness of the results with respect to the demand distribution.

Next, by comparing the above two cases, we conclude that both the manufacturer and the whole supply chain always prefer retailer-adopting case. The retailer, however, prefers manufacturer-adopting 3D printing case, for most problem parameters.

Finally, we acknowledge that, as a preliminary study of 3D printing in the context of supply chain, this research has several limitations. Addressing these limitations can lead to valuable future research. On the quality side, researchers can allow for lower revenue for 3D printed products since their quality is not yet the same as traditionally manufactured products. Addressing this question adds another level of complexity to our problem. To simplify the problem, researchers can assume 3D printing capacity is exogenous. On the demand side, researchers can utilize price dependent demand models to optimize over the selling price. However, they should be mindful that in such a case, the retailer can ask for a premium because 3D printed products can be customized to each customer. Moreover, researchers may also want to consider the positioning of the 3D printing technology. They can study the case where both the manufacturer and the retailer have the option of adopting 3D printing technology. The proposed model in Chapter 4 can be extend to such a scenario.

Bibliography

Agrawal, N., and Smith, S., 2017. Optimal inventory management using retail prepacks. *Working paper*, Santa Clara University, Santa Clara, CA.

APC (American Power Conversion), 2005. Determining total cost of ownership for data center and network room infrastructure. White paper, #6, http://www.apc.com, retrieved on May 2, 2018.

Arikan, E., 2014 and Jammernegg W.. 2014 The single period inventory model under dual sourcing and productcarbon footprint constraint. *Int. J. Production Economics*, 157, pp. 15-23.

Baker, K. R., 1989. Lot-sizing procedures and a standard data set: a reconciliation of the literature. *Journal of Manufacturing and Operations Management*, 2(3), pp. 199-221.

Bansal S, Transchel S, 2014. LManaging supply risk for vertically differentiated co-products. *Production Operations Management*, 23(9), pp. 1577-1598.

Banciu M. and P. Mirchandani, 2013, Technical Note: New Results Concerning Probability Distributions with Increasing Generalized Failure Rates, *Operations Research*, 61 (4), pp. 925-931.

Bernstein F. and A. G. Kök, 2009, Dynamic Cost Reduction through Process Improvement in assembly Networks, *Management Science*, 55(4), pp. 552-567.

Boyabatl O, Kleindorfer PR, Koontz SR 2011. Integrating long-term and short-term contracting in beef supply chains. *Management Science*, 57(10), pp. 1771-1787.

Cachon G., 2003. Supply Chain Coordination with Contracts. *Handbooks in Operations Research and Management Science*, 11, pp. 227-339.

Cachon G. and Larivier M., 2005. Supply Chain Coordination with Revenue-Sharing Contracts: Strengths and Limitations. *Management Science*, 5 (1), pp. 30-44.

Chen L. and Cui Y. and Lee H., 2017. Retailing with 3D Printing, note = SSRN Scholarly Paper ID 3031566.

Chen, S., Lee, H., and Moinzadeh, K., 2018. Pricing schemes in cloud computing: Utilization-based versus reservation-based. *Production and Operations Management*. First published online on 15 May, 2018.

Chen, S., Moinzadeh, K., and Tan, Y., 2018. On-demand and low-priority services with limited capacity: Pricing schemes for cloud platforms. Working paper, University of Washington, Seattle, WA.

Chen Y, Tomlin B, Wang Y, 2017. Dual Coproduct Technologies: Implications for Process Development and Adoption. *Manufacturing and Service Operations Management* , 19(4), pp. 692-712.

Cisco Inc., 2016. Cisco global cloud index: forecast and methodology 2015-2020. White Paper, Cisco Corp.

Cohen M., R. Lobel and G. Perakis, 2016, The Impact of Demand Uncertainty on Consumer Subsidies for Green Technology Adoption, *Management Science*, 62(5), pp. 1235-1258.

Comez N. and E. Stecke and M. akanyldrm, 2012, In-Season Transshipments Among Competitive Retailers, *Manufacturing and Service Operations Management*, 14(2), pp 290-300.

Conner P. and G. Manogharan and A. Martof and L. Rodomsky and C. Rodomsky and D. Jordan and J. Limperos, Making Sense of 3-D Printing: Creating a Map of Additive Manufacturing Products and Services, *Additive Manufacturing*, 1-4, 64-76.

Chen K. and Yang L., 2014. Random yield and coordination mechanisms of a supply chain with emergency backup sourcing. *International Journal of Production Research*, 52(16).

Dieker, A., Ghosh, S., and Squillante, M. S., 2016. Optimal resource capacity management for stochastic networks. *Operations Research*, 65(1), pp. 221-241.

Dignan, L., 2015. 35 percent of cloud computing spending is wasted, says RightScale. *ZDNet*, http://www.zdnet.com/article/35-percent-of-cloud-computing-spending-is-wasted-says-rightscale/.

Dillon T., Wu C., Change E., 2008. Cloud Computing: Issues and Challenges. *IEEE International Conference on Advanced Information Networking and Applications.*

Dobson G. and C. Yano, Product offering, priding, and make-to-stock/make-to-order Decision with Shared Capacity, Production and Operations Management Society, 11, 3, 2002

Dong L, Kouvelis P, Wu X , 2014. The value of operational flexibility in the presence of input and output price uncertainties with oil refining applications. *Management Science*, 60(12), pp. 2908-2926.

Dong L. and D. Shi and F. Zhang, 2017, 3D Printing vs. Traditional Flexible Technology: Implications for Operations Strategy, Available at SSRN: https://ssrn.com/abstract=2847731.

Dong L. and N. Rudi, Who Benefits from transshipment? Exogenous vs. Endogenous Wholesale Prices, 2004, *Managementn Science*, 50(5), pp. 648-657.

Eppen, G. D., Martin, R. K., and Schrage, L., 1989. OR practice – a scenario approach to capacity planning. *Operations Research*, 37(4), pp. 517-527.

Erlenkotter, D., 1974. Note – A dynamic programming approach to capacity expansion with specialization. *Management Science*, 21(3), pp. 360-362.

Fisher, M., Ramdas, K., and Zheng, Y. S., 2001. Ending inventory valuation in multi-period production scheduling. *Management Science*, 47(5), pp. 679-692.

Fong, C. O. and Rao, M. R., 1975. Capacity expansion with two producing regions and concave costs. *Management Science*, 22(3), pp. 331-339.

Frye, 2013. Infographic: the life cycle of a server. *TechRepublic*, https://www.techrepublic.com/blog/data-center/infographic-the-life-cycle-of-a-server/, accessed on November 5, 2017.

Galante, G. and Bona, L.C.E.D., 2012. A survey on cloud computing elasticity. In *Proceedings of the 2012 IEEE/ACM Fifth International Conference on Utility and Cloud Computing*, pp. 263-270.

Gartner Press Release, 2017. Gartner says worldwide public cloud services market to grow 18 Percent in 2017. https://www.gartner.com/newsroom/id/3616417, Gartner, Inc.

Ge Z. and Q. Hu and Y. Xia, 2014, Firms' Research and Developement Cooperation Behavior in a Supply Chain, *Production and Operations Management*, 23(4), pp. 599-609.

Goyal M. and S. Netessine, 2007, Strategic Technology Choice and Capacity Investment under Demand Uncertainty, *Management Science*, 53(2), pp. 192-207.

Goyal, S. K., 1974. Determination of optimum packaging frequency of items jointly replenished. *Management Science*, 21(4), pp. 436-443.

Gross D., 2013, Obama's Speech Highlights Rise of 3D Printing, http://www.cnn.com/2013/02/13/tech/innovation/obama-3d-printing/, accessed 1-Feb-2017.

Grunewald S., 2016, Is Disney Getting into the 3D Printing Business? They Just Filed for Three New Related Patents, https://3dprint.com/123998/disney-3d-printing-patents, accessed 1-Feb-2017.

Grunewald S., 2016 3D Printing Healthcare Market is Expected to Grow by 18% Annually Until 2020, https://3dprint.com/137461/3d-printing-healthcare-market/, accessed 1-Feb-2017.

Gupta S., 2008, Research Note - Channel Structure with Knowledge Spillovers, *Marketing Science*, 27(2), pp. 247-261.

Hall N., 2016, France launches funding initiative for 3D printing, https://3dprintingindustry.com/news/france-launches-funding-initiative-3d-printing-81141/, Online; accessed 1-Feb-2017

Harms R. and Yamartino M., 2010, THE ECONOMICS OF THE CLOUD, *Microsoft White Paper*.

Hohn M. S., 2010, Relational Supply Contracts, *Springer*, Berlin.

Ishii A., Cooperative Research and Development between Vertically Related Firms with Spillovers, *International Journal of Industrial Organization*, 22(8), pp. 1213-1235.

Janakiraman G. and S. Seshadri, 2017 Dual Sourcing Inventory Systems: On Optimal Policiesand the Value of Costless Returns, *Production and Operations Management Society*, 26 (2), pp. 203-210

Kepes, 2015. 30% of servers are sitting "comatose" according to research. *Forbes.com*, https: //www.forbes.com/sites/benkepes/2015/06/03/30-of-servers-are-sitting-comatose-according-to-research/#5cd4c69859c7, accessed on November 5, 2017.

Khouja, M. and Goyal, S., 2008. A review of the joint replenishment problem literature: 1989-2005. *European Journal of Operational Research*, 186 (1), pp. 1-16.

Kilcioglu, C., and Maglaras, C., 2015. Revenue maximization for cloud computing services. *SIGMETRICS Performance Evaluation Review*, 43(3), 76.

Lariviere M. and E. Porteus, 2001, Selling to the Newsvendor: An Analysis of Price-only Contracts, *Manufacturing and Service Operations Management*, 3(4), pp. 293-305.

Li, B. and Kumar, S., 2018. Should you kill or embrace your competitor: Cloud service and competition strategy. *Production and Operations Management*, 27(5), 822-838.

Li, S. and Tirupati, D., 1994. Dynamic capacity expansion problem with multiple products: Technology selection and timing of capacity additions. *Operations Research*, 42(5), pp. 958-976.

Li H. and Y. Wang and R. Yin and T. J. Kull and T. Y. Choi, 2012, Target Pricing: Demand-Side versus Supply-Side Approaches, *Journal of Production Economics*, 136(1), pp. 172-184.

Liang C. and S. P. Sethi and R. Shi and J. Zhang, 2014, Inventory Sharing with Transshipment: Impacts of Demand Distribution Shapes and Setup Costs, *Production and Operations Management Society*, 23 (10), pp. 1779-1794

Licata J., 2013, How 3D printing could revolutionise the solar energy industry, https://www.theguardian.com/environment/blog/2013/feb/22/3d-printing-solar-energy-industry, accessed 1-Feb-2017.

Lu, L. and Qi, X., 2011. Dynamic lot sizing for multiple products with a new joint replenishment model. *European Journal of Operational Research*, 212(1), pp. 74-80.

Lu, Y., Song, J. S., and Yao, D. D., 2003. Order fill rate, leadtime variability, and advance demand information in an assemble-to-order system. *Operations Research*, 51(2), pp. 292-308.

Lu, Y., Song, J. S., and Zhao, Y., 2010. No-holdback allocation rules for continuous-time assemble-to-order systems. *Operations Research*, 58(3), pp. 691-705.

Luss, H., 1982. Operations research and capacity expansion problems: A survey. *Operations Research*, 30(5), pp. 907-947.

Luss, H., 1984. Capacity expansion planning for a single facility product line. *European Journal of Operational Research*, 18(1), pp. 27-34.

Manne, A.S., 1961. Capacity expansion and probabilistic growth. *Econometrica*, pp. 632-649.

Martínez-Costa, C., Mas-Machuca, M., Benedito, E., and Corominas, A., 2014. A review of mathematical programming models for strategic capacity planning in manufacturing. *International Journal of Production Economics*, 153, pp. 66-85.

Mell, P. and Grance, T., 2011. The NIST definition of cloud computing: recommendations of the National Institute of Standards and Technology. *NIST Special Publication 800-145*.

Ng T, Fowler J, MoK I , 2012. Robust demand service achievement for the co-production newsvendor. *IIE Transaction*, 44, 327,341.

Pasternack B., 1985, Optimal Pricing and Return Policies for Perishable Commodities, *Marketing Science*, 31(8), pp. 166-176.

Van Mieghem J., 1998, Investment Strategies for Flexible Resources, *Management Science*, 44(8), pp. 1071-1078.

Mishra, A. K., Hellerstein, J. L., Cirne, W., and Das, C. R., 2010. Towards characterizing cloud backend workloads: insights from Google compute clusters. *ACM SIGMETRICS Performance Evaluation Review*, 37(4), pp. 34-41.

Nahmias S, Moinzadeh K ,1997. Lot sizing with randomly graded yields. *Operations Research*, 47(6), pp. 974-986.

Nimrodi, J., 2014. 10 facts that you didn't know about server farms. *Cloudyn*, https:// www.cloudyn.com /blog/10-facts-didnt-know-server-farms/.

Porteus E., 1985, Investing in Reduced Setups in the EOQ Model, *Management Science*, 31(8), pp. 998-1010.

Protalinski, E., 2018. Microsoft reports $30.1 billion in Q4 2018 revenue: Azure up 89%, Surface up 25%, and Windows up 7%. *VentureBeat.com*. July 19, 2018.

Ramasesh R. and J. Keith Ord and J. C. Hayya and A. Pan, 1991, Sole Versus Dual Sourcing in Stochastic Lead-Time (s, Q) Inventory Models, *Management Schience*, 22 (11), pp. 428-443.

Rotolo D. and D. Hicks and B. Marlin, 2015, What Is an Emerging Technology?, *Research Policy*, 44 (10), pp. 1827-1843.

Sethuraman N. and A. Parlakturk and J. Swaminathan, 2018, Personal Fabrication as an Operational Strategy: Value of Delegating Production to Customer, SSRN Scholarly Paper ID 3170011

Shen, S., van Beek, V., and Iosup, A., 2015. Statistical characterization of business-critical workloads hosted in cloud datacenters. In *Cluster, Cloud and Grid Computing (CCGrid), 2015 15th IEEE/ACM International Symposium on* (pp. 465-474), IEEE.

Silver, E.A., 1976. A simple method of determining order quantities in joint replenishments under deterministic demand. *Management Science*, 22(12), pp. 1351-1361.

Silver, E. A. and Meal, H. C., 1973. A heuristic for selecting lot size quantities for the case of a deterministic time-varying demand rate and discrete opportunities for replenishment. *Production and Inventory Management*, 14(2), pp. 64-74.

Song, J. S., 1998. On the order fill rate in a multi-item, base-stock inventory system. *Operations Research*, 46(6), pp. 831-845.

Song, J. S., 2002. Order-based backorders and their implications in multi-item inventory systems. *Management Science*, 48(4), pp. 499-516.

Song J. and Y. Zhang, 2016, Stock or Print? Impact of 3D Printing on Spare Parts Logistics, Available at SSRN: https://ssrn.com/abstract=2884459.

Fine C. and R. Freund, 1990, Optimal Investment in Product-Flexible Manufacturing Capacity, *Management Science*, 36(4), pp. 449-466.

Song, J. S. and Zipkin, P., 2003. Supply chain operations: Assemble-to-order systems. *Handbooks in Operations Research and Management Science*, 11, pp. 561-596.

Stadtler, H., 2000. Improved rolling schedules for the dynamic single-level lot-sizing problem. *Management Science*, 46(2), pp. 318-326.

Taylor T., 2002. upply Chain Coordination Under Channel Rebates with Sales Effort Effects. *Management Science*, 48(8), pp. 992-1007.

Tomlin B, Wang Y, 2008 . Pricing and operational recourse in coproduction systems. *Management Science*, 54(3), pp. 522-537.

Toosi, A. N., Vanmechelen, K., Khodadadi, F., and Buyya, R., 2016. An auction mechanism for cloud spot markets. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 11(1), 2.

Tsay A., 1999 . The Quantity Flexibility Contract and Supplier-Customer Incentives. *Management Science*, 45(10), pp. 1339-1358.

Van Den Heuvel, W. and Wagelmans, A. P., 2005. A comparison of methods for lot-sizing in a rolling horizon environment. *Operations Research Letters*, 33(5), pp. 486-496.

Van, H. N., Tran, F. D., and Menaud, J. M., 2009. SLA-aware virtual resource management for cloud infrastructures. *9th IEEE International Conference on Computer and Information Technology*, pp. 1-8.

Van Mieghem, J. A., 2003. Commissioned paper: Capacity management, investment, and hedging: Review and recent developments. *Manufacturing & Service Operations Management*, 5(4), pp.269-302.

Veeraraghavan S. and A. Scheller-Wolf, Now or Later: A Simple Policy for Effective Dual Sourcing in Capacitated Systems, 2008, *Operations Research*, 56(4), pp. 850-864.

Wagner, H. M. and Whitin, T. M., 1958. Dynamic version of the economic lot size model. *Management Science*, 5(1), pp. 89-96.

Wang T. and Hu Q., 2010. Coordination of Supply Chain with Advertise-Setting Newsvendor. *IEEE Xplore*, pp. 391-395.

Wang Y. and W. Gilland and B. Tomlin. Mitigating Supply Risk: Dual Sourcing or Process Improvement?. *Manufacturing and Service Operations Management*, 12(3), pp. 489-510.

Westerweel B. and R. Basten, 2016, Traditional or Additive Manufacturing? Assessing component design options through lifecycle cost analysis.

Westerweel B. and R. Basten, 2017, Traditional or Additive Manufacturing? Assessing component design options through lifecycle cost analysis.

Westerweel B. and R. Basten and G. Van Houtum, 2018, Printing spare parts at remote locations: Fulfilling the promise of additive manufacturing, Working paper

Yu H. and A. Zeng and L. Zhao, 2009 . Single or dual sourcing: decision-making in the presence of supply chain disruption risks. *The International Journal of Management SCIENCE*, 37, pp. 788-800.

Xu, H., and Li, B., 2013. Dynamic cloud pricing for revenue maximization. *IEEE Transactions on Cloud Computing*, 1(2), 158-171.

Zipkin, P. H., 2000. *Foundations of inventory management*. McGraw-Hill Higher Education.

Appendix

Proof. **Lemma 3.1.** The total fixed expansion cost is obvious. We have shown how to calculate the total replenishment (purchase) cost when cluster 1 is the leading cluster. It is straightforward to obtain this cost when cluster 2 is the leading cluster by swapping the indices of cluster 1 and cluster 2. Thus, (3.3) can be seen as a generalization of both cases when cluster $l \in \{1, 2\}$ is the leading cluster.

We now analyze the total holding cost. Our analysis below is based on the case where cluster 1 is the leading cluster, as the result can be easily generalized when cluster $l \in \{1, 2\}$ is the leading cluster.

For attribute 2, the holding of inventory during the cycle $[T_n, T_{n+1}]$ is calculated as follows:

$$
\begin{aligned}
&\sum_{k=1}^{m} \int_{\tau_{k,n}}^{\tau_{k+1,n}} [a_{21}Q_{k,1,n} - D_2(\tau_{k,n}, t)]dt + \int_{\tau_{m+1,n}}^{T_{n+1}} [a_{22}Q_{m+1,2,n} - D_2(\tau_{m+1,n}, t)]dt \\
&= \sum_{k=1}^{m} (\tau_{k+1,n} - \tau_{k,n}) D_2(\tau_{k,n}, \tau_{k+1,n}) - \sum_{k=1}^{m} \int_{\tau_{k,n}}^{\tau_{k+1,n}} D_2(\tau_{k,n}, t)dt \\
&\quad + (T_{n+1} - \tau_{m+1,n}) D_2(\tau_{m+1,n}, T_{n+1}) - \int_{\tau_{m+1,n}}^{T_{n+1}} D_2(\tau_{m+1,n}, t)dt \\
&= \sum_{k=1}^{m} (\tau_{k+1,n} - \tau_{k,n}) D_2(\tau_{k,n}, \tau_{k+1,n}) - \sum_{k=1}^{m} \int_{\tau_{k,n}}^{\tau_{k+1,n}} [D_2(T_n, t) - D_2(T_n, \tau_{k,n})]dt \\
&\quad + (T_{n+1} - \tau_{m+1,n}) D_2(\tau_{m+1,n}, T_{n+1}) - \int_{\tau_{m+1,n}}^{T_{n+1}} D_2(\tau_{m+1,n}, t)dt \\
&= \sum_{k=1}^{m} (\tau_{k+1,n} - \tau_{k,n}) D_2(\tau_{k,n}, \tau_{k+1,n}) - \int_{T_n}^{\tau_{m+1,n}} D_2(T_n, t)dt + \sum_{k=1}^{m} (\tau_{k+1,n} - \tau_{k,n}) D_2(T_n, \tau_{k,n}) \\
&\quad + (T_{n+1} - \tau_{m+1,n}) D_2(\tau_{m+1,n}, T_{n+1}) - \int_{\tau_{m+1,n}}^{T_{n+1}} D_2(\tau_{m+1,n}, t)dt \\
&= \sum_{k=1}^{m} (\tau_{k+1,n} - \tau_{k,n}) D_2(T_n, \tau_{k+1,n}) - \int_{T_n}^{\tau_{m+1,n}} D_2(T_n, t)dt \\
&\quad + (T_{n+1} - \tau_{m+1,n}) [D_2(T_n, T_{n+1}) - D_2(T_n, \tau_{m+1,n})] - \int_{\tau_{m+1,n}}^{T_{n+1}} [D_2(T_n, t) - D_2(T_n, \tau_{m+1,n})]dt \\
&= -F_2(T_n, T_{n+1}) + \sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) D_2(T_n, \tau_{k+1,n}).
\end{aligned}
$$

The first equality of the above calculation is due to (3.1).

For attribute 1, the holding of inventory during the cycle $[T_n, T_{n+1}]$ is calculated as follows. First, the inventory of attribute 1 held during $[\tau_{k,n}, \tau_{k+1,n}]$ is given by

$$\begin{aligned}
&\int_{\tau_{k,n}}^{\tau_{k+1,n}} \left[a_{11} \sum_{l=1}^{k} Q_{l,1,n} - D_1(T_n, \tau_{k,n}) - D_1(\tau_{k,n}, t) \right] dt \\
&= (\tau_{k+1,n} - \tau_{k,n}) \left[\frac{a_{11}}{a_{21}} D_2(T_n, \tau_{k+1,n}) - D_1(T_n, \tau_{k,n}) \right] - \int_{\tau_{k,n}}^{\tau_{k+1,n}} [D_1(T_n, t) - D_1(T_n, \tau_{k,n})] dt \\
&= (\tau_{k+1,n} - \tau_{k,n}) \frac{a_{11}}{a_{21}} D_2(T_n, \tau_{k+1,n}) - \int_{\tau_{k,n}}^{\tau_{k+1,n}} D_1(T_n, t) dt.
\end{aligned}$$

Next, for the last replenishment until the end of this cycle, the inventory of attribute 1 held is given by

$$\begin{aligned}
&\int_{\tau_{m+1,n}}^{T_{n+1}} \left[a_{11} \sum_{k=1}^{m} Q_{k,1,n} + a_{12} Q_{m+1,2,n} - D_1(T_n, \tau_{m+1,n}) - D_1(\tau_{m+1,n}, t) \right] dt \\
&= \int_{\tau_{m+1,n}}^{T_{n+1}} \left[\frac{a_{11}}{a_{21}} \sum_{k=1}^{m} D_2(\tau_{k,n}, \tau_{k+1,n}) + a_{12} Q_{m+1,2,n} - D_1(T_n, \tau_{m+1,n}) - D_1(\tau_{m+1,n}, t) \right] dt \\
&= \int_{\tau_{m+1,n}}^{T_{n+1}} \left[\frac{a_{11}}{a_{21}} D_2(T_n, \tau_{m+1,n}) + a_{12} Q_{m+1,2,n} - D_1(T_n, \tau_{m+1,n}) - D_1(\tau_{m+1,n}, t) \right] dt.
\end{aligned}$$

But since $a_{22} Q_{m+1,2,n} = D_2(\tau_{m+1,n}, T_{n+1})$ due to (3.1), the above expression can be reduced to

$$(T_{n+1} - \tau_{m+1,n}) \left[\frac{a_{11}}{a_{21}} D_2(T_n, T_{n+1}) - (\frac{a_{11}}{a_{21}} - \frac{a_{12}}{a_{22}}) D_2(\tau_{m+1,n}, T_{n+1}) \right] - \int_{\tau_{m+1,n}}^{T_{n+1}} D_1(T_n, t) dt.$$

Hence, the total inventory of attribute 1 held during the cycle $[T_n, T_{n+1}]$ will be

$$\sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) \frac{a_{11}}{a_{21}} D_2(T_n, \tau_{k+1,n}) - (T_{n+1} - \tau_{m+1,n}) (\frac{a_{11}}{a_{21}} - \frac{a_{12}}{a_{22}}) D_2(\tau_{m+1,n}, T_{n+1}) - F_1(T_n, T_{n+1}).$$

Therefore, the total holding cost for this cycle will be:

$$\sum_{i=1}^{2} \frac{h_i a_{i1}}{a_{21}} \left[\sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) D_2(T_n, \tau_{k+1,n}) \right]$$
$$-h_1(T_{n+1} - \tau_{m+1,n})(\frac{a_{11}}{a_{21}} - \frac{a_{12}}{a_{22}}) D_2(\tau_{m+1,n}, T_{n+1}) - \sum_{i=1}^{2} h_i F_i(T_n, T_{n+1}).$$

Since $D_2(\tau_{m+1,n}, T_{n+1}) = a_{22} Q_{m+1,2,n} = \alpha a_{22}[a_{11} D_2(T_n, T_{n+1}) - a_{21} D_1(T_n, T_{n+1})]$ and $w_1 = h_1 a_{11} + h_2 a_{21}$, the total holding cost can also be written as follows:

$$\sum_{i=1}^{2} \frac{h_i a_{i1}}{a_{21}} \left[\sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) D_2(T_n, \tau_{k+1,n}) \right]$$
$$-h_1(T_{n+1} - \tau_{m+1,n}) \big(\frac{a_{11}}{a_{21}} D_2(T_n, T_{n+1}) - D_1(T_n, T_{n+1}) \big) - \sum_{i=1}^{2} h_i F_i(T_n, T_{n+1})$$
$$= \sum_{k=1}^{m} (\tau_{k+1,n} - \tau_{k,n}) \frac{w_1 D_2(T_n, \tau_{k+1,n})}{a_{21}} + \sum_{i=1}^{2} h_i [(T_{n+1} - \tau_{m+1,n}) D_i(T_n, T_{n+1}) - F_i(T_n, T_{n+1})]$$
$$= \sum_{k=1}^{m} (\tau_{k+1,n} - \tau_{k,n}) \frac{w_1 D_2(T_n, \tau_{k+1,n})}{a_{21}} + \psi_n.$$

□

Proof. **Lemma 3.2.** The calculation of the total purchase and fixed expansion costs are straightforward, so here we focus on that of the total holding cost. According to (3.5), the

total holding cost of attribute i is

$$
\begin{aligned}
& h_i \left\{ \sum_{k=1}^{m+1} \int_{\tau_{k,n}}^{\tau_{k+1,n}} [a_{i1} Q_{k,1,n} + a_{i2} Q_{k,2,n} - D_i(\tau_{k,n}, t)] dt \right\} \\
= \; & h_i \left\{ \sum_{k=1}^{m+1} \int_{\tau_{k,n}}^{\tau_{k+1,n}} [D_i(\tau_{k,n}, \tau_{k+1,n}) - D_i(\tau_{k,n}, t)] dt \right\} \\
= \; & h_i \left\{ \sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) D_i(\tau_{k,n}, \tau_{k+1,n}) - \sum_{k=1}^{m+1} \int_{\tau_{k,n}}^{\tau_{k+1,n}} [D_i(T_n, t) - D_i(T_n, \tau_{k,n})] dt \right\} \\
= \; & h_i \left\{ \sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) D_i(\tau_{k,n}, \tau_{k+1,n}) - \int_{T_n}^{T_{n+1}} D_i(T_n, t) dt + \sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) D_i(T_n, \tau_{k,n}) \right\} \\
= \; & h_i \left\{ \sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) D_i(T_n, \tau_{k+1,n}) - \int_{T_n}^{T_{n+1}} D_i(T_n, t) dt \right\} \\
= \; & h_i \left\{ \sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) D_i(T_n, \tau_{k+1,n}) - F_i(T_n, T_{n+1}) \right\}.
\end{aligned}
$$

Therefore, the total holding cost of the two attributes is given by (3.7). □

Proof. **Proposition 3.1.** The proof consists of two steps. First, we will show that at one extreme, $\beta = 1$, the optimal total cost of the simultaneous-replenishment policy is strictly lower than that of the sequential-replenishment policy. Second, we will show that at the other extreme, $\beta = 2$, the optimal total cost of the simultaneous-replenishment policy is strictly higher than that of the sequential-replenishment policy. Then, since the optimal total cost of the simultaneous-replenishment policy is increasing in β, the result of this proposition must hold.

Case 1: when $\beta = 1$.

In that case, the total fixed expansion and purchase costs of the simultaneous- and sequential-replenishment policies are the same, so we can focus on comparing the total holding costs of the two policies. Specifically, we will show that for any given sequential-replenishment policy, we can find a simultaneous-replenishment policy with a lower total

holding cost. Then, the optimal total cost of the simultaneous-replenishment policy must be smaller than that of the sequential-replenishment policy. The proof is as follows:

For any sequential-replenishment policy, according to (3.4), the total holding cost is equal to

$$\begin{aligned} & h_f \left[\sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) D_f(T_n, \tau_{k+1,n}) - F_f(T_n, T_{n+1}) \right] \\ + \quad & h_l \left[\sum_{k=1}^{m} (\tau_{k+1,n} - \tau_{k,n}) \frac{a_{ll}}{a_{fl}} D_f(T_n, \tau_{k+1,n}) + (T_{n+1} - \tau_{m+1,n}) D_l(T_n, T_{n+1}) - F_l(T_n, T_{n+1}) \right]. \end{aligned}$$

By contrast, with the same expansion points (i.e., the $\tau_{k,n}$'s), the total holding cost of the simultaneous-replenishment policy is, according to (3.7),

$$\begin{aligned} & h_f \left[\sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) D_f(T_n, \tau_{k+1,n}) - F_f(T_n, T_{n+1}) \right] \\ + \quad & h_l \left[\sum_{k=1}^{m} (\tau_{k+1,n} - \tau_{k,n}) D_l(T_n, \tau_{k+1,n}) + (T_{n+1} - \tau_{m+1,n}) D_l(T_n, T_{n+1}) - F_l(T_n, T_{n+1}) \right]. \end{aligned}$$

Note that, according to Assumption 3.1, $a_{ll}/a_{fl} > D_l(T_n, \tau_{k+1,n})/D_f(T_n, \tau_{k+1,n})$. Then, (1) is strictly smaller than (1), which completes the proof for Case 1.

Case 1: when $\beta = 2$.

In that case, we will show that for any given simultaneous-replenishment policy, we can find a sequential-replenishment policy with the same total purchase and expansion costs but with a lower total holding cost. Then, the optimal total cost of the sequential-replenishment policy must be smaller than that of the simultaneous-replenishment policy. The proof is as follows:

For any simultaneous-replenishment policy, the total holding cost between any two con-

secutive expansion points ($\tau_{k,n}$ and $\tau_{k+1,n}$) is, according to (3.5),

$$h_1 \int_{\tau_{k,n}}^{\tau_{k+1,n}} [a_{11}Q_{k,1,n}+a_{12}Q_{k,2,n}-D_1(\tau_{k,n},t)]dt+h_2 \int_{\tau_{k,n}}^{\tau_{k+1,n}} [a_{21}Q_{k,1,n}+a_{22}Q_{k,2,n}-D_2(\tau_{k,n},t)]dt. \tag{1}$$

By contrast, consider a specifc sequential-replenishment policy, which results in leveled capacities of both attributes at the expansion times of the given simultaneous-replenishment policy, as follows. One needs to purchase $Q_{k,1,n}$ units of cluster 1 at expansion time $\tau_{k,n}$ and then $Q_{k,2,n}$ units of cluster 2 at a specific time point, $y_{k,n}$, when the excess capacity of attribute 2 reaches zero. In such a way, the excess capacities of the attributes will both reach zero at $\tau_{k+1,n}$, as they do so under the simultaneous-replenishment policy with the same purchase quantities. It is obvious that $\tau_{k,n} < y_{k,n} < \tau_{k+1,n}$.

The purchase and fixed expansion costs of the given simultaneous-replenishment policy and the constructed sequential-replenishment policy are the same. Then, we can focus on comparing the total holding costs of these two policies. Specifically, the total holding cost of the sequential-replenishment policy is

$$\begin{aligned} & h_1 \left\{ \int_{\tau_{k,n}}^{y_{k,n}} [a_{11}Q_{k,1,n} - D_1(\tau_{k,n},t)]dt + \int_{y_{k,n}}^{\tau_{k+1,n}} [a_{11}Q_{k,1,n} + a_{12}Q_{k,2,n} - D_1(\tau_{k,n},t)]dt \right\} \\ + \quad & h_2 \left\{ \int_{\tau_{k,n}}^{y_{k,n}} [a_{21}Q_{k,1,n} - D_2(\tau_{k,n},t)]dt + \int_{y_{k,n}}^{\tau_{k+1,n}} [a_{22}Q_{k,2,n} - D_2(y_{k,n},t)]dt \right\}. \end{aligned} \tag{2}$$

Note that $a_{21}Q_{k,1,n} = D_2(\tau_{k,n}, y_{k,n})$ and $a_{22}Q_{k,2,n} = D_2(y_{k,n}, \tau_{k+1,n})$. Then,

$$h_2 \int_{y_{k,n}}^{\tau_{k+1,n}} [a_{22}Q_{k,2,n} - D_2(y_{k,n},t)]dt = h_2 \int_{y_{k,n}}^{\tau_{k+1,n}} [a_{21}Q_{k,1,n} + a_{22}Q_{k,2,n} - D_2(\tau_{k,n},t)]dt.$$

Therefore, comparing (2) to (1), the total holding cost of the constructed sequential-replenishment policy is strictly smaller than that of the given simultaneous-replenishment

policy, and the difference is $h_1 \sum_{i=1}^{2}(y_{k,n} - \tau_{k,n})a_{i2}Q_{k,2,n}$, which completes the proof for Case 2. □

Proof. **Lemma 3.3.** Note that $\tau_{m+1,n}$ should satisfy equation (3.8):

$$\frac{a_{11}}{a_{21}}D_2(T_n, \tau_{m+1,n}) + \frac{a_{12}}{a_{22}}D_2(\tau_{m+1,n}, T_{n+1}) = D_1(T_n, T_{n+1}).$$

That is,

$$\frac{a_{11}}{a_{21}}(e^{\lambda_2\tau_{m+1,n}} - e^{\lambda_2 T_n}) + \frac{a_{12}}{a_{22}}D_2(e^{\lambda_2 T_{n+1}} - e^{\lambda_2\tau_{m+1,n}}) = e^{\lambda_1 T_{n+1}} - e^{\lambda_1 T_n}.$$

Then,

$$\tau_{m+1,n} = \frac{1}{\lambda_2}\ln\left(\frac{a_{21}(a_{22}e^{\lambda_1 T_{n+1}} - a_{12}e^{\lambda_2 T_{n+1}}) + a_{22}(a_{11}e^{\lambda_2 T_n} - a_{21}e^{\lambda_1 T_n})}{\alpha}\right) = T_n + \frac{\ln\chi_n}{\lambda_2},$$

where χ_n is as stated in this lemma. □

Proof. **Proposition 3.2.** Since $t_n = (\tau_{m+1,n} - T_n)/m = \ln\chi_n/(\lambda_2 m)$,

$$D_2(0, kt_n) = \gamma\big(e^{\lambda_2 k t_n} - 1\big) = \gamma(\chi_n^{\frac{k}{m}} - 1).$$

Thus, the objective function, (3.10), will be

$$A(m+1) + \frac{w_1 e^{\lambda_2 T_n}}{a_{21}}\frac{\ln\chi_n}{\lambda_2 m}\sum_{k=1}^{m}\gamma(\chi_n^{\frac{k}{m}} - 1) + \psi_n = A(m+1) + \frac{\gamma w_1 e^{\lambda_2 T_n}}{a_{21}\lambda_2}\left(\frac{\sum_{k=1}^{m}\chi_n^{\frac{k}{m}}}{m} - 1\right)\ln\chi_n + \psi_n.$$

Furthermore, to prove that the above objective function is convex in m, we only need to prove that the second term of the objective function is convex in m, as the first term is

linear in m and the third term is independent of m. Note that

$$\frac{\gamma w_1 e^{\lambda_2 T_n}}{a_{21}\lambda_2}\left(\frac{\sum_{k=1}^{m}\chi_n^{\frac{k}{m}}}{m}-1\right)\ln\chi_n=\frac{\gamma w_1 e^{\lambda_2 T_n}}{a_{21}\lambda_2}\left(\frac{\chi_n^{\frac{1}{m}}(\chi_n-1)}{m(\chi_n^{\frac{1}{m}}-1)}-1\right)\ln\chi_n.$$

Next, to prove the above function is convex in m, we need to prove the following function is convex: define

$$Z(m)=\frac{\chi_n^{\frac{1}{m}}(\chi_n-1)}{m(\chi_n^{\frac{1}{m}}-1)}.$$

Then,

$$\frac{d^2Z(m)}{dm^2}=\frac{2m^2\left(\chi_n^{\frac{1}{m}}-1\right)^2-4m\left(\chi_n^{\frac{1}{m}}-1\right)\ln\chi_n+\left(\chi_n^{\frac{1}{m}}+1\right)(\ln\chi_n)^2}{m^5\left(\chi_n^{\frac{1}{m}}-1\right)^3}.$$

It is easy to show that $\chi_n>1$. Then, the numerator of the above expression is greater than $2\left(m\left(\chi_n^{\frac{1}{m}}-1\right)-\ln\chi_n\right)^2$, which is greater than zero. The denumerator is also greater than zero. Hence, the second-order derivative of $Z(m)$ with respect to m is positive, implying that $Z(m)$ is convex in m.

Next, consider the EB. Since $D_2(T_n,\tau_{k+1})=(k/m)D_2(T_n,\tau_{m+1,n})$ for any $k=1,\ldots,m$, that is

$$e^{\lambda_2(\tau_{k+1}-T_n)}-1=\frac{k}{m}(e^{\lambda_2(\tau_{m+1,n}-T_n)}-1)=\frac{k}{m}(\chi_n-1).$$

Thus,

$$\tau_{k+1,n}=T_n+\frac{1}{\lambda_2}\ln\left[\frac{k}{m}\chi_n+(1-\frac{k}{m})\right].$$

Then,

$$\tau_{m+1,n}-\sum_{k=1}^{m}\frac{\tau_{k,n}}{m}=\frac{\ln\chi_n}{\lambda_2}-\frac{1}{\lambda_2 m}\sum_{k=1}^{m-1}\ln\left[\frac{k}{m}\chi_n+(1-\frac{k}{m})\right]$$

Moreover, $D_2(T_n, \tau_{m+1,n}) = \gamma e^{\lambda_2 T_n}(e^{\lambda_2(\tau_{m+1,n}-T_n)} - 1) = \gamma e^{\lambda_2 T_n}(\chi_n - 1)$. Therefore, the objective function (i.e., (**??**)) reduces to

$$A(m+1) + \frac{\gamma w_1 e^{\lambda_2 T_n}}{\lambda_2 a_{21}}(\chi_n - 1)\left[\ln \chi_n - \frac{\sum_{k=1}^{m-1} \ln\left(\frac{k}{m}\chi_n + (1 - \frac{k}{m})\right)}{m}\right] + \psi_n.$$

□

Proof. **Proposition 3.3.** Substituting $D_2(T_n, \tau_{k+1,n}) = \lambda_{2n} k t_n$, where $t_n = (\tau_{m+1,n} - T_n)/m$, into the objective function (3.10), the objective function becomes:

$$A(m+1) + \frac{w_1 \lambda_{2n}}{2a_{21}}(\tau_{m+1,n} - T_n)^2\big(1 + \frac{1}{m}\big) + \psi_n. \tag{3}$$

The objective function is convex in m. Thus, the optimal (real) value of m should satisfy the first-order condition, which is

$$\begin{aligned} \widehat{m} &= \sqrt{\frac{w_1\lambda_{2n}}{2Aa_{21}}}(\tau_{m+1,n} - T_n) \\ &= \sqrt{\frac{w_1\lambda_{2n}}{2Aa_{21}}}\frac{a_{21}(\lambda_{1n}a_{22} - \lambda_{2n}a_{12})(T_{n+1} - T_n)}{\alpha\lambda_{2n}} = \sqrt{\frac{w_1 a_{21}}{2A\lambda_{2n}}}\frac{(\lambda_{1n}a_{22} - \lambda_{2n}a_{12})(T_{n+1} - T_n)}{\alpha} \end{aligned} \tag{4}$$

The second equation above is due to (3.8), where we use the approximation $D_2(T_n, \tau_{m+1,n}) = \lambda_{2n}(\tau_{m+1,n} - T_n)$, $D_2(\tau_{m+1,n}, T_{n+1}) = \lambda_{2n}(T_{n+1} - \tau_{m+1,n})$, and $D_2(T_n, T_{n+1}) = \lambda_{2n}(T_{n+1} - T_n)$. □

Proof. **Proposition 3.4.** Substituting $D_i(T_n, T_n + kt_n) = \gamma e^{\lambda_i T_n}\left(e^{\lambda_i k t_n} - 1\right)$ and $t_n =$

$(T_{n+1} - T_n)/(m+1)$ into the objective function (3.15), we obtain

$$\begin{aligned} & t_n \sum_{i=1}^{2} \left\{ h_i \sum_{k=1}^{m+1} D_i(T_n, T_n + kt_n) \right\} \\ = \quad & t_n \sum_{i=1}^{2} \left\{ h_i \sum_{k=1}^{m+1} \gamma e^{\lambda_i T_n} \left(e^{\lambda_i k t_n} - 1 \right) \right\} = \frac{T_{n+1} - T_n}{m+1} \sum_{i=1}^{2} \left\{ h_i \gamma e^{\lambda_i T_n} \sum_{k=1}^{m+1} \left(e^{\lambda_i k \frac{T_{n+1} - T_n}{m+1}} - 1 \right) \right\}, \end{aligned}$$

which gives (3.17). Furthermore, (3.17) is a convex function of m, noticing the similarity between (3.12) and (3.17). That is, the same proof for the convexity of (3.12) applies to that of (3.17).

Next, note that the total cost under simultaneous-replenishment policy is:

$$\beta A(m+1) + \sum_{i=1}^{2} h_i \left[\sum_{k=1}^{m+1} (\tau_{k+1,n} - \tau_{k,n}) D_i(T_n, \tau_{k+1,n}) + F_i(T_n, T_{n+1}) \right]$$

Furthermore, $D_i(T_n, \tau_{k+1,n}) = \frac{k}{m} D_i(T_n, T_{n+1})$. *Thus,*

$$\tau_{(k+1,n)} = T_n + \frac{1}{\lambda_i} ln(\frac{k}{m}(e^{(\lambda_i (T_{n+1} - T_n))} - 1) + 1)$$

Therefore, total cost simplifies to:

$$
\begin{aligned}
& A(m+1)+\sum_{i=1}^{2} h_i\left[\sum_{k=1}^{m+1}(\tau_{k+1,n}-\tau_{k,n})\frac{k}{m}D_i(T_n,T_{n+1})+F_i(T_n,T_{n+1})\right] \\
= \;& A(m+1)+\sum_{i=1}^{2}\frac{h_iD_i(T_n,T_{n+1})}{m}\left[\sum_{k=1}^{m+1}(\tau_{k+1,n}-\tau_{k,n})k+F_i(T_n,T_{n+1})\right] \\
= \;& A(m+1)+\sum_{i=1}^{2}\frac{h_iD_i(T_n,T_{n+1})}{m}\left[(m+1)T_{n+1}-\sum_{k=1}^{m+1}\tau_{k,n}+F_i(T_n,T_{n+1})\right] \\
= \;& A(m+1)+\sum_{i=1}^{2}\frac{h_iD_i(T_n,T_{n+1})}{m}\left[\sum_{k=0}^{m}(T_{n+1}-\tau_{k+1,n})+F_i(T_n,T_{n+1})\right] \\
= \;& A(m+1)+\sum_{i=1}^{2}\frac{h_iD_i(T_n,T_{n+1})}{m}\left[\sum_{k=0}^{m}\left(T_{n+1}-(T_n+\frac{1}{\lambda_i}ln(\frac{k}{m}(e^{(\lambda_i(T_{n+1}-T_n))}-1)+1))\right)\right] \\
= \;& A(m+1)+\sum_{i=1}^{2}\frac{h_ie^{\lambda_iT_n}(e^{\lambda_i(t)}-1)}{m}\left[\sum_{k=0}^{m}\left(t-\frac{1}{\lambda_i}ln\left(\frac{k}{m}e^{t\lambda_i}+(1-\frac{k}{m})\right)\right)+F_i(T_n,T_{n+1})\right]
\end{aligned}
$$

□

Proof. **Proposition 3.5.** We substitute $D_i(T_n, T_n + kt_n) = \lambda_{in}kt_n$ into (3.15) and obtain

$$
\begin{aligned}
& t_n\sum_{i=1}^{2}\left\{h_i\sum_{k=1}^{m+1}D_i(T_n,T_n+kt_n)\right\} \\
= \;& t_n\sum_{i=1}^{2}\left\{h_i\sum_{k=1}^{m+1}\lambda_{in}kt_n\right\}=t_n^2\frac{(m+2)(m+1)}{2}\sum_{i=1}^{2}h_i\lambda_{in}=\frac{(T_{n+1}-T_n)^2(m+2)}{2(m+1)}\sum_{i=1}^{2}h_i\lambda_{in}.
\end{aligned}
$$

That is, the objective function (3.15) becomes:

$$\beta A(m+1)+\frac{(T_{n+1}-T_n)^2(m+2)}{2(m+1)}\sum_{i=1}^{2}h_i\lambda_{in}+\phi_n.$$

Then, the first-order condition with respect to m is as follows:

$$\beta A - \left(\frac{(T_{n+1} - T_n)^2 \sum_{i=1}^{2} h_i \lambda_{in}}{2} \right) \frac{1}{(m+1)^2} = 0,$$

which gives (3.19). Furthermore, the second-order condition is satisfied, i.e., the objective function is convex. □

Proof. **Proposition 3.6.** Using the approximate approach, we can obtain an closed-form expression for the optimal total cost. That is, substituting (4) into the objective function (3), we obtain

$$\begin{aligned} & A(m+1) + \frac{w_1 \lambda_{2n}}{2a_{21}} (\tau_{m+1,n} - T_n)^2 \big(1 + \frac{1}{m}\big) \\ &= A + \sqrt{\frac{2Aw_1 a_{21}}{\lambda_{2n}}} \frac{\lambda_{1n} a_{22} - \lambda_{2n} a_{12}}{\alpha} (T_{n+1} - T_n) + \frac{w_1}{2a_{21}\lambda_{2n}} \left[\frac{a_{22}(\lambda_{2n} a_{11} - \lambda_{1n} a_{21})}{\alpha} \right]^2 (T_{n+1} - T_n)^2 \end{aligned}$$

And ψ_n is equal to

$$\sum_{i=1}^{2} h_i [(T_{n+1} - \tau_{m+1,n}) D_i(T_n, T_{n+1}) - F_i(T_n, T_{n+1})] = (T_{n+1} - T_n)^2 \sum_{i=1}^{2} h_i \lambda_{in} \left[\frac{a_{22}(\lambda_{2n} a_{11} - \lambda_{1n} a_{21})}{\alpha \lambda_{2n}} - \frac{1}{2} \right].$$

Thus, the optimal total cost (including the holding and the fixed expansion costs) is equal to

$$TC(T_n, T_{n+1}) = A + \sqrt{\frac{2Aw_1 a_{21}}{\lambda_{2n}}} \frac{\lambda_{1n} a_{22} - \lambda_{2n} a_{12}}{\alpha} (T_{n+1} - T_n) + \Phi_n (T_{n+1} - T_n)^2,$$

where Φ_n is independent of T_n and T_{n+1} and is equal to

$$\Phi_n = \frac{(a_{11}\lambda_{2n} - a_{21}\lambda_{1n})[h_1(a_{12}^2 a_{21}\lambda_{2n} + a_{11}a_{22}^2\lambda_{1n} - 2a_{12}a_{21}a_{22}\lambda_{1n}) + h_2 a_{22}^2(\lambda_{2n}a_{11} - \lambda_{1n}a_{21})]}{2\alpha^2 \lambda_{2n}}.$$

Then, the total cost rate of the cycle is as follows:

$$\frac{TC(T_n, T_{n+1})}{T_{n+1} - T_n} = \frac{A}{T_{n+1} - T_n} + \Phi_n(T_{n+1} - T_n) + \sqrt{\frac{2Aw_1a_{21}}{\lambda_{2n}}}\frac{(\lambda_{1n}a_{22} - \lambda_{2n}a_{12})}{\alpha}.$$

The second-order condition of the total cost rate with respect to T_{n+1} is

$$\frac{2A}{(T_{n+1} - T_n)^3} > 0$$

Therefore, the total cost rate is a strictly convex function of T_{n+1}, and T^*_{n+1} can be found by solving the first-order condition. That is, $T^*_{n+1} - T_n = \sqrt{A/\Phi_n}$. □

Proof. **Proposition 4.2**. If $v = 0$, then the equilibrium is provided in Corollary 4.1, which gives the manufacturer a profit of $(w_b - c_m)q_{nv}(w_b)$. If $v = 1$, by substituting out $q_m = q_{nv}(w) - q_p$, we can write Problem (4.3) as

$$\begin{aligned} \pi^{3D}_M = \max_{n,q_p,w\geq 0} \quad & (w - c_m)q_{nv}(w) + (c_m - c_p)q_p - (nK + S) \\ \text{s.t.} \quad & q_p \leq nQ \leq q_{nv}(w) \\ & \pi^b_R(w) \geq A, \end{aligned}$$

If $c_m \leq c_p$, then $q_p = 0$ and 3D printing is not adopted. Otherwise, sinec the objective function is non-decreasing in q_p, we set $q_p = nQ$, and the problem simplifies further to

$$\begin{aligned} \pi^{3D}_M = \max_{n,w\geq 0} \quad & (w - c_m)q_{nv}(w) + \left(c_m - c_p - \tfrac{K}{Q}\right)nQ - S \\ \text{s.t.} \quad & nQ \leq q_{nv}(w) \\ & \pi^b_R(w) \geq A. \end{aligned}$$

If $c_m < c_p + \frac{K}{Q}$, the objective is decreasing in n, we set $n = 0$, and the problem reduces to the benchmark problem with no 3D printers, whose solution is given in Corollary 4.1.

Otherwise, the objective is non-decreasing in n and, due to the first constraint, we set $n = \frac{q_{nv(w)}}{Q}$. The model simplifies to

$$\begin{aligned} \pi_M^{3D} = \max_{w \geq 0} \quad & \left(w - c_p - \tfrac{K}{Q}\right) q_{nv}(w) - S \\ \text{s.t.} \quad & \pi_R^b(w) \geq A, \end{aligned}$$

whose solution is provided in Corollary 4.2: the optimal wholesale price is $\min\{\hat{w}_s, w_m^A\}$. Finally, the 3D option is preferred by the manufacturer over the traditional manufacturing option if and only if

$$\left(\min\{\hat{w}_s, w_m^A\} - c_p - \frac{K}{Q}\right) q_{nv}(\min\{\hat{w}_s, w_m^A\}) - S \geq \pi_M^b. \; \square \tag{5}$$

□

Proof. **Proposition 4.3**. Since $r > w_p + c_p$, the objective function is increasing in $E[q_p(D)]$. Therefore, the second constraint is tight, and, letting $R \triangleq r - w_p - c_p$, the objective can be written as

$$\begin{aligned} \pi(q_m, n) &\triangleq E[r\min\{q_m, D\}] - w_m q_m + E[R\min\{nQ, \max\{D - q_m, 0\}\}] - nK \\ &= \int_0^{q_m} rxf(x)dx + \int_{q_m}^{U} rq_m f(x)dx - w_m q_m + \int_{q_m}^{q_m+nQ} R(x - q_m)f(x)dx + \int_{q_m+nQ}^{U} RnQf(x)dx - nK \\ &= \int_0^{q_m} rxf(x)dx + rq_m(1 - F(q_m)) - w_m q_m + \int_{q_m}^{q_m+nQ} R(x - q_m)f(x)dx + RnQ(1 - F(q_m + nQ)) \\ &\quad -nK. \end{aligned}$$

The gradient and hessian of the objective, with respect to q_m and n, are

$$\begin{pmatrix} r - (w_p + c_p)F(q_m) - w_m - RF(q_m + nQ) \\ RQ(1 - F(q_m + nQ)) - K \end{pmatrix}$$

and

$$\begin{pmatrix} -(w_p+c_p)f(q_m)-Rf(q_m+nQ) & -RQf(q_m+nQ) \\ -RQf(q_m+nQ) & -RQ^2f(q_m+nQ) \end{pmatrix},$$

respectively. A straightforward analysis shows that all the first principal minors of the Hessian matrix are negative and the second principal minor is nonnegative; therefore, the objective function is concave. The first order condition reduces to the following two cases:

1. $K > RQ$: In this case, since $\frac{\partial \pi(q_m,n)}{\partial n} < 0$ for all q_m, we conclude $n^* = 0$, which implies $q_m^* = F^{-1}\left(1-\frac{w_m}{r}\right)$.

2. $K \leq RQ$: If $q_m \geq F^{-1}\left(1-\frac{K}{RQ}\right)$, then $\frac{\partial(\pi(q_m,n))}{\partial n} \leq 0$, and subsequently $n^* = 0$. However, if $q_m \leq F^{-1}\left(1-\frac{K}{RQ}\right)$, then $\frac{\partial(\pi(q_m,n))}{\partial n} = 0$ has a unique root, namely $n(q_m) = \frac{1}{Q}(F^{-1}\left(1-\frac{K}{RQ}\right) - q_m)$. Therefore,

$$n^*(q_m) = \begin{cases} \frac{1}{Q}\left(F^{-1}\left(1-\frac{K}{RQ}\right)-q_m\right), & q_m \leq F^{-1}\left(1-\frac{K}{RQ}\right) \\ 0, & q_m \geq F^{-1}\left(1-\frac{K}{RQ}\right). \end{cases} \tag{6}$$

Plugging Equation (6) back into the objective, we obtain a continuous univariate function of q_m:

$$\pi(q_m) \triangleq \begin{cases} \int_0^{q_m} rxf(x)dx + rq_m(1-F(q_m)) - w_mq_m + \int_{q_m}^{F^{-1}\left(1-\frac{K}{RQ}\right)} R(x-q_m)f(x)dx, & q_m \leq F^{-1}\left(1-\frac{K}{RQ}\right) \\ \int_0^{q_m} rxf(x)dx + rq_m(1-F(q_m)) - w_mq_m, & q_m \geq F^{-1}\left(1-\frac{K}{RQ}\right). \end{cases} \tag{7}$$

To obtain the first case, note that at $n(q_m) = \frac{1}{Q}(F^{-1}\left(1-\frac{K}{RQ}\right) - q_m)$, the expression $RnQ(1-F(q_m+nQ)) - nK$ simplifies to zero. We refer to $F^{-1}\left(1-\frac{K}{RQ}\right)$ as the break point, and we define $R_1 = \{q_m|\ q_m \leq F^{-1}\left(1-\frac{K}{RQ}\right)\}$, and $R_2 = \{q_m|\ q_m >$

$F^{-1}\left(1-\frac{K}{RQ}\right)\}$. The first and second derivatives of $\pi(q_m)$ are

$$\frac{\partial(\pi(q_m))}{\partial q_m} = \begin{cases} (w_p+c_p)(1-F(q_m)) - (w_m - \frac{K}{Q}), & q_m \leq F^{-1}\left(1-\frac{K}{RQ}\right) \\ r(1-F(q_m)) - w_m, & q_m \geq F^{-1}\left(1-\frac{K}{RQ}\right) \end{cases} \tag{8}$$

and

$$\frac{\partial^2(\pi(q_m))}{\partial q_m^2} = \begin{cases} -(w_p+c_p)f(q_m), & q_m \leq F^{-1}\left(1-\frac{K}{RQ}\right) \\ -rf(q_m), & q_m > F^{-1}\left(1-\frac{K}{RQ}\right), \end{cases} \tag{9}$$

respectively. Note that the first derivative is continuous, and the objective function is strictly concave. Due to the strict concavity of the objective, the first order condition will result in at most one root in $R_1 \cup R_2$. Next, based on the sign of the first derivative at the break point,

$$\frac{\partial(\pi(q_m)}{\partial q_m}\Big|_{F^{-1}\left(1-\frac{K}{RQ}\right)} = \frac{rK}{RQ} - w_m,$$

we partition Case 2 into two subcases.

(a) $\frac{rK}{RQ} > w_m$: In this case, $\frac{\partial(\pi_R^{3D}(q_m))}{\partial q_m}\Big|_{F^{-1}\left(1-\frac{K}{RQ}\right)} > 0$. Therefore, by strict concavity, any root must occur in R_2, and we see that $q_m^* = F^{-1}\left(1-\frac{w_m}{r}\right)$ is the unique root, since $w_m \leq r$. Since the root is in R_2, $n^* = 0$.

(b) $\frac{rK}{RQ} \leq w_m$: In this case, $\frac{\partial(\pi_R^{3D}(q_m))}{\partial q_m}\Big|_{F^{-1}\left(1-\frac{K}{RQ}\right)} \leq 0$. Again, by strict concavity, any root must occur in R_1. If $w_m \leq w_p + c_p + \frac{K}{Q}$, then $q_m^* = F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)$ is the unique root (note that, since $r > R \triangleq (r-w_p-c_p)$, $\frac{rK}{RQ} \leq w_m$ implies that $w_m \geq \frac{K}{Q}$) and, since the root is in R_1, $n^* = \frac{1}{Q}\left(F^{-1}\left(1-\frac{K}{RQ}\right) - F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)\right)$. If $w_m > w_p + c_p + \frac{K}{Q}$, $\frac{\partial(\pi(q_m)}{\partial q_m} \leq 0$ for all q_m, and we conclude that $q_m^* = 0$; also, since the root is in R_1, $n^* = \frac{1}{Q}F^{-1}\left(1-\frac{K}{RQ}\right)$.

The above analysis can be summarized as

$$(q_m^*, n^*) = \begin{cases} \left(F^{-1}\left(1-\frac{w_m}{r}\right), 0\right), & \frac{w_m}{r}R < \frac{K}{Q} \\ \left(F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right), \frac{1}{Q}\left(F^{-1}\left(1-\frac{K}{RQ}\right) - F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)\right)\right., & \frac{K}{Q} \leq \frac{w_m}{r}R \text{ and } w_m - w_p - c_p \leq \frac{K}{Q} \\ (0, \frac{1}{Q}(F^{-1}\left(1-\frac{K}{RQ}\right))), & \frac{K}{Q} \leq \frac{w_m}{r}R \text{ and } w_m - w_p - c_p > \frac{K}{Q}; \end{cases} \tag{10}$$

note that straightforward algebra shows that $w_m - w_p - c_p \leq \frac{w_m}{r}R$, which results in the conditions stated in the proposition's statement. □ □

Proof. **Proposition 4.4**. We obtain the optimal profits by plugging (10) back into (7), and we utilize the fact that $\pi_R^b(r,w)$ can be written as $\int_0^{F^{-1}\left(1-\frac{w}{r}\right)} rxf(x)dx$. The first case of (10), plugged into the second case of (7), simplifies to the Newsvendor problem, and $\pi_R^{3D}(w_m, w_p) = \pi_R^b(r, w_m)$. The second and third cases of (10) are plugged into the first case of (7). In the second case, we obtain

$$\begin{aligned} \pi_R^{3D}(w_m, w_p) &= \int_0^{F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)} rxf(x)dx + r\frac{w_m-\frac{K}{Q}}{w_p+c_p}F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right) - w_m F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right) \\ &\quad + \int_{F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)}^{F^{-1}\left(1-\frac{K}{RQ}\right)} Rxf(x)dx - \int_{F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)}^{F^{-1}\left(1-\frac{K}{RQ}\right)} RF^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right) f(x)dx \\ &= \int_0^{F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)} rxf(x)dx + \int_{F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)}^{F^{-1}\left(1-\frac{K}{RQ}\right)} Rxf(x)dx + X, \end{aligned}$$

where

$$X = r\frac{w_m-\frac{K}{Q}}{w_p+c_p}F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right) - w_m F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right) - \int_{F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)}^{F^{-1}\left(1-\frac{K}{RQ}\right)} RF^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right) f(x)dx.$$

We next show that $X = 0$. The last term in X can be rewritten as

$$\int_{F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)}^{F^{-1}\left(1-\frac{K}{RQ}\right)} RF^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)f(x)dx = RF^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)\left(\frac{w_m-\frac{K}{Q}}{w_p+c_p}-\frac{K}{RQ}\right),$$

which implies

$$X = F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)\left(r\frac{w_m-\frac{K}{Q}}{w_p+c_p}-w_m-R\frac{w_m-\frac{K}{Q}}{w_p+c_p}+\frac{K}{Q}\right)=0,$$

since $r - R = w_p + c_p$. This simplification implies that

$$\begin{aligned}
\pi_R^{3D}(w_m, w_p) &= \int_0^{F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)} rxf(x)dx + \int_{F^{-1}\left(1-\frac{w_m-\frac{K}{Q}}{w_p+c_p}\right)}^{F^{-1}\left(1-\frac{K}{RQ}\right)} Rxf(x)dx \\
&= \frac{r}{w_p+c_p}\pi_R^b\left(w_p+c_p, w_m-\frac{K}{Q}\right)+\pi_R^b\left(R,\frac{K}{Q}\right)-\frac{R}{w_p+c_p}\pi_R^b\left(w_p+c_p, w_m-\frac{K}{Q}\right) \\
&= \pi_R^b\left(w_p+c_p, w_m-\frac{K}{Q}\right)+\pi_R^b\left(R,\frac{K}{Q}\right).
\end{aligned}$$

In the third case, $\pi_R^{3D}(w_m, w_p) = \int_0^{F^{-1}\left(1-\frac{K}{RQ}\right)} Rxf(x)dx = \pi_R^b\left(R, \frac{K}{Q}\right)$, which completes the proof. □ □

Proof. **Lemma 4.1**. π_M^1 can be written as

$$\begin{aligned}
\max_{w_m, w_p \geq 0} \quad & (w_m - c_m)\left(1-\tfrac{w_m}{r}\right)U \\
\text{s.t.} \quad & \tfrac{w_m}{r}(r - w_p - c_p) < \tfrac{K}{Q} \\
& \pi_R^b(r, w_m) \geq 0,
\end{aligned}$$

since, from Proposition 4.3, $q_m^* = F^{-1}\left(1-\frac{w_m}{r}\right) = \left(1-\frac{w_m}{r}\right)U$ for a uniform distribution on $[0, U]$, and the π_R^{3D} expression is from Proposition 4.4. Since w_p does not affect the objective function, by setting it to $r - c_p \geq 0$, we can get the largest solution space on w_m (because

$\frac{K}{Q}$ is positive), which eliminates the first constraint. For the moment ignoring the constraint $\pi_R^b(r, w_m) \geq 0$, optimizing the unconstrained concave objective results in $w_m^* = \frac{r+c_m}{2}$; since $w_m^* \leq r$, the constraint $\pi_R^b(r, w_m) \geq 0$ is satisfied. The equilibrium order quantity is therefore $q_m^* = \left(1 - \frac{w_m^*}{r}\right) U = \left(\frac{r-c_m}{2r}\right) U$, and the resulting retailer profit in equilibrium is

$$\begin{aligned} \pi_R^{3D}(w_m^*, w_p^*) &= r \int_0^{q_m^*} x f(x) dx + r \int_{q_m^*}^{U} q_m^* f(x) dx - c_m q_m^* \\ &= \frac{r}{2U}(q_M^*)^2 + \frac{r}{U} q_m^* (U - q_m^*) - c_m q_m^* \\ &= \frac{3}{2}\left(\frac{r - c_m}{2r}\right)^2 U r. \end{aligned}$$

Finally, the manufacturer's profit is $\pi_M^1 = (w_m^* - c_m)\left(1 - \frac{w_m^*}{r}\right) U = \frac{1}{r}(\frac{r-c_m}{2})^2 U$ □ □

Proof. **Lemma 4.2**. Using Propositions 4.3 and 4.4, π_M^2 can be written as

$$\begin{aligned} \max_{w_m, w_p \geq 0} \quad & (w_m - c_m)\left(1 - \tfrac{w_m - \frac{K}{Q}}{w_p + c_p}\right) U + w_p E[q_p^*(D)] \cdot \mathbb{1}\{n^* > 0\} \\ \text{s.t.} \quad & w_m - w_p - c_p \leq \tfrac{K}{Q} \\ & \tfrac{K}{Q} \leq \tfrac{w_m}{r}(r - w_p - c_p) \\ & q_p^*(D) = \min\left\{\left(\left(1 - \tfrac{K}{(r - w_p - c_p)Q}\right) - \left(1 - \tfrac{w_m - \frac{K}{Q}}{w_p + c_p}\right)\right) U, \max\left\{0, D - \left(1 - \tfrac{w_m - \frac{K}{Q}}{w_p + c_p}\right) U\right\}\right\} \\ & \pi_R^b\left(w_p + c_p, w_m - \tfrac{K}{Q}\right) + \pi_R^b\left(r - w_p - c_p, \tfrac{K}{Q}\right) \geq 0. \end{aligned}$$

For $(w_m, w_p) \in P_2$, $w_m - w_p - c_p \leq \frac{K}{Q} \Leftrightarrow w_m - \frac{K}{Q} \leq w_p + c_p \Rightarrow \pi_R^b\left(w_p + c_p, w_m - \frac{K}{Q}\right) \geq 0$. Similarly, for $(w_m, w_p) \in P_2$, $\frac{K}{Q} \leq \frac{w_m}{r}(r - w_p - c_p) \Rightarrow \frac{K}{Q} \leq r - w_p - c_p \Rightarrow \pi_R^b\left(r - w_p - c_p, \frac{K}{Q}\right) \geq 0$. Thus, for $(w_m, w_p) \in P_2$, the constraint $\pi_R^b\left(w_p + c_p, w_m - \frac{K}{Q}\right) + \pi_R^b\left(r - w_p - c_p, \frac{K}{Q}\right) \geq 0$ is satisfied.

We make the following change of variables: $x = \left(1 - \frac{w_m - \frac{K}{Q}}{w_p + c_p}\right) U$ and $y = \left(1 - \frac{K}{(r - w_p - c_p)Q}\right) U$, with inverses $w_p = r - c_p - \frac{K}{Q(1 - y/U)}$ and $w_m = r(1 - x/U) - \frac{K(1 - x/U)}{Q(1 - y/U)} + \frac{K}{Q}$. The constraints

in P_2 translate to $0 \leq x \leq y \leq U$. Thus, π_M^2 can be written as

$$\max_{0\leq x\leq y\leq U} \left(\frac{r}{U}(U-x) - \frac{K(U-x)}{Q(U-y)} + \frac{K}{Q} - c_m\right) x + w_p E[\min\{y-x, \max\{0, D-x\}\}].$$

The expected amount of 3D printed products can be calculated as

$$\begin{aligned} E[q_p^*(D)] &= \int_0^U \min\{y-x, \max\{0,\xi-x\}\} \frac{1}{U} d\xi \\ &= \frac{1}{U}\left(\int_0^x \min\{y-x, 0\} d\xi + \int_x^U \min\{y-x, \xi-x\} d\xi\right) \\ &= \frac{1}{U}\left(\int_x^y (\xi - x) d\xi + \int_y^U (y-x) d\xi\right) \\ &= \frac{1}{2U}(x^2 - y^2) + (y-x). \end{aligned}$$

We may again rewrite π_M^2 as

$$\max_{0\leq x\leq y\leq U} \left(\frac{r}{U}(U-x) - \frac{K(U-x)}{Q(U-y)} + \frac{K}{Q} - c_m\right) x + \left(r - c_p - \frac{KU}{Q(U-y)}\right)\left(\frac{1}{2U}(x^2-y^2) + y - x\right). \tag{11}$$

We define the objective function as

$$\Gamma(x,y) \triangleq \left(\frac{r}{U}(U-x) - \frac{K(U-x)}{Q(U-y)} + \frac{K}{Q} - c_m\right) x + \left(r - c_p - \frac{KU}{Q(U-y)}\right)\left(\frac{1}{2U}(x^2-y^2) + y - x\right)$$

. We first fix y and find the optimal $x^*(y)$, which is a function of y; we then reinsert $x^*(y)$ into the objective function, and solve for the optimal y^*. To begin, we determine the partial derivative of $\Gamma(x,y)$ with respect to x:

$$\frac{\partial \Gamma(x,y)}{\partial x} = \left(\frac{K}{Q(U-y)} - \frac{r}{U} - \frac{c_p}{U}\right) x + \frac{K}{Q} - c_m + c_p = \mathcal{A}x + \mathcal{B}, \tag{12}$$

where $\mathcal{A} \triangleq \frac{K}{Q(U-y)} - \frac{r}{U} - \frac{c_p}{U}$ and $\mathcal{B} \triangleq \frac{K}{Q} - c_m + c_p$. Note that the first partial derivative with

respect to x is linear in x. Recall that we require $0 \leq x \leq y$. We consider the following four cases:

1. If $\mathcal{B} \geq 0$ and $\mathcal{A}y + \mathcal{B} \geq 0$, the derivative is always positive, implying that the objective function is increasing in x, and consequently $x^*(y) = y$.

2. If $\mathcal{B} \geq 0$ and $\mathcal{A}y + \mathcal{B} < 0$, the objective is strictly concave, and there is a unique maximizer at $x^*(y) = -\mathcal{B}/\mathcal{A}$; note that $\mathcal{B} \geq 0$ and $\mathcal{A}y + \mathcal{B} < 0$ imply that $\mathcal{A} < 0$, which further implies that $0 \leq -\mathcal{B}/\mathcal{A} < y$.

3. If $\mathcal{B} < 0$ and $\mathcal{A}y + \mathcal{B} < 0$, the derivative is always negative, implying that the objective function is decreasing in x, and consequently $x^*(y) = 0$.

4. If $\mathcal{B} < 0$ and $\mathcal{A}y + \mathcal{B} \geq 0$, the objective function is convex, so either $x = 0$ or $x = y$ returns the maximum. We first see that

$$\begin{aligned}
\Gamma(0,y) - \Gamma(y,y) &= \left(r - c_p - \frac{KU}{Q(U-y)}\right)\left(y - \frac{y^2}{2U}\right) - y\left(\frac{r}{U}(U-y) - c_m\right) \\
&= y\left(\left(r - c_p - \frac{KU}{Q(U-y)}\right)\left(1 - \frac{y}{2U}\right) - \left(\frac{r}{U}(U-y) - c_m\right)\right) \\
&= y\left(\frac{Ky}{2Q(U-y)} - \frac{KU}{Q(U-y)} + \frac{(r+c_p)y}{2U} + c_m - c_p\right) \\
&= \frac{y}{U-y}\left(\frac{Ky}{2Q} - \frac{KU}{Q} + \left(\frac{(r+c_p)y}{2U} + c_m - c_p\right)(U-y)\right) \\
&= \frac{y}{U-y}\gamma(y),
\end{aligned}$$

where $\gamma(y) \triangleq \frac{Ky}{2Q} - \frac{KU}{Q} + \left(\frac{(r+c_p)y}{2U} + c_m - c_p\right)(U-y)$. We can write $\gamma(y)$ as a quadratic in y, as follows:

$$\gamma(y) = -\frac{(r+c_p)}{2U}y^2 + \left(\frac{K}{2Q} + \frac{(r+c_p)}{2} - c_m + c_p\right)y + \left(c_m - c_p - \frac{K}{Q}\right)U.$$

Since $\mathcal{B} < 0 \Leftrightarrow c_m > c_p + \frac{K}{Q}$, $\gamma(0) > 0$, and there is one positive root and one negative root. The positive root of this quadratic is

$$\hat{y} \triangleq \frac{U}{(r+c_p)}\left(\frac{K}{2Q} + \frac{(r+c_p)}{2} - c_m + c_p + \sqrt{\left(\frac{K}{2Q} + \frac{(r+c_p)}{2} - c_m + c_p\right)^2 + 2(r+c_p)\left(c_m - c_p - \frac{K}{Q}\right)}\right);$$

since $\gamma(U) = -\frac{KU}{2Q} < 0$, we may conclude that $\hat{y} < U$. Therefore, the concave quadratic $\gamma(y)$ is non-negative for $y \in [0, \hat{y}]$. Thus,

$$x^*(y) = \begin{cases} 0, & y \in [0, \hat{y}] \\ y, & y \in (\hat{y}, U] \end{cases}$$

maximizes $\Gamma(x, y)$ in this case.

We now summarize the above four cases:

$$x^*(y) = \begin{cases} 0, & \{\mathcal{B} < 0 \text{ and } \mathcal{A}y + \mathcal{B} < 0\} \text{ or } \{\mathcal{B} < 0 \text{ and } \mathcal{A}y + \mathcal{B} \geq 0 \text{ and } y \in [0, \hat{y}]\} \\ -\frac{\mathcal{B}}{\mathcal{A}}, & \mathcal{B} \geq 0 \text{ and } \mathcal{A}y + \mathcal{B} < 0 \\ y, & \{\mathcal{B} \geq 0 \text{ and } \mathcal{A}y + \mathcal{B} \geq 0\} \text{ or } \{\mathcal{B} < 0 \text{ and } \mathcal{A}y + \mathcal{B} \geq 0 \text{ and } y \in (\hat{y}, U]\}. \end{cases} \tag{13}$$

We next simplify the conditions in Equation (13). Letting $L(y) \triangleq \mathcal{A}y + \mathcal{B} = \left(\frac{K}{Q(U-y)} - \frac{r}{U} - \frac{c_p}{U}\right) y + \frac{K}{Q} - c_m + c_p$, we derive simpler conditions on y that are equivalent to $L(y) \geq 0$:

$$\begin{aligned} & \left(\frac{K}{Q(U-y)} - \frac{(r+c_p)}{U}\right) y + \frac{K}{Q} - c_m + c_p \geq 0 \\ \Leftrightarrow \quad & \frac{1}{U-y}\left(\left(\frac{K}{Q} - \frac{(r+c_p)(U-y)}{U}\right) y + \frac{K}{Q}(U-y) - (c_m - c_p)(U-y)\right) \geq 0 \\ \Leftrightarrow \quad & \frac{\delta(y)}{U-y} \geq 0, \end{aligned}$$

where $\delta(y) \triangleq \frac{(r+c_p)}{U}y^2 + (c_m - r - 2c_p)y + \left(\frac{K}{Q} - c_m + c_p\right)U$. The discriminant of the convex quadratic $\delta(y)$ is $\Delta \triangleq (r + 2c_p - c_m)^2 - 4(r + c_p)\left(\frac{K}{Q} - c_m + c_p\right) = (r + c_m)^2 - 4(r + c_p)\frac{K}{Q}$. Therefore, if $\Delta < 0$, then $L(y) \geq 0$ for all y; otherwise $L(y) \geq 0$ for $y \leq \xi_1 \triangleq \frac{(r+2c_p-c_m-\sqrt{\Delta})U}{2(r+c_p)}$ or $y \geq \xi_2 \triangleq \frac{(r+2c_p-c_m+\sqrt{\Delta})U}{2(r+c_p)}$, where ξ_1 and ξ_2 are the roots of $\delta(y)$. Note that the minimizer of $\delta(y)$ is $\tilde{y} \triangleq \left(\frac{(r+2c_p-c_m)}{2(r+c_p)}\right)U \in [0, U]$; note also that $\delta(U) = \frac{KU}{Q} > 0$, which implies that $\xi_2 \in [\tilde{y}, U]$. If $\mathcal{B} \geq 0$, then $\gamma(0) = \mathcal{B}U \geq 0$, and we conclude that $\xi_1 \in [0, \tilde{y}]$; otherwise, $\xi_1 < 0$.

We next reinsert $x^*(y)$ from Equation (13), rewritten more simply using the values of (Δ, ξ_1, ξ_2), into the objective function, obtaining $\Gamma(x^*(y), y)$, which we then maximize over y. We consider two cases, depending on the value of the discriminant $\Delta = (r+c_m)^2 - 4(r+c_p)\frac{K}{Q}$ (which only depends on problem data):

1. If $(r + c_m)^2 - 4(r + c_p)\frac{K}{Q} < 0$, then $\mathcal{A}y + \mathcal{B} \geq 0$ is always true, and Equation (13) can be written as

$$x^*(y) = \begin{cases} 0, & \mathcal{B} < 0 \text{ and } y \in [0, \hat{y}] \\ y, & \mathcal{B} \geq 0 \text{ or } \{\mathcal{B} < 0 \text{ and } y \in (\hat{y}, U]\}. \end{cases}$$

   Note that $(r + c_m)^2 - 4(r + c_p)\frac{K}{Q} < 0$ implies that $\mathcal{B} = \frac{K}{Q} - c_m + c_p > 0$: for a contradiction, assume that $\frac{K}{Q} \leq c_m - c_p$. Then, $0 > (r + c_m)^2 - 4(r + c_p)\frac{K}{Q} \geq (r + c_m)^2 - 4(r + c_p)(c_m - c_p) = (r - c_m)^2 + 4c_p(r - c_m) + 4c_p^2 \geq 0$. Therefore, Equation (13) can be written as $x^*(y) = y$, for all $y \in [0, U]$, and

$$\Gamma(y, y) = \left(\frac{r}{U}(U - y) - c_m\right)y.$$

   This concave function is maximized at $y^* = \left(\frac{r-c_m}{2r}\right)U$.

2. If $(r + c_m)^2 - 4(r + c_p)\frac{K}{Q} \geq 0$, then $\mathcal{A}y + \mathcal{B} \geq 0 \Leftrightarrow \{y \leq \xi_1 \text{ or } y \geq \xi_2\}$, and Equation

(13) can be written as

$$x^*(y) = \begin{cases} 0, & \{\mathcal{B} < 0 \text{ and } \xi_1 < y < \xi_2\} \text{ or } \{\mathcal{B} < 0 \text{ and } \{y \le \xi_1 \text{ or } y \ge \xi_2\} \text{ and } y \in [0, \hat{y}]\} \\ -\frac{\mathcal{B}}{\mathcal{A}}, & \mathcal{B} \ge 0 \text{ and } \xi_1 < y < \xi_2 \\ y, & \{\mathcal{B} \ge 0 \text{ and } \{y \le \xi_1 \text{ or } y \ge \xi_2\}\} \text{ or } \{\mathcal{B} < 0 \text{ and } \{y \le \xi_1 \text{ or } y \ge \xi_2\} \text{ and } y \in (\hat{y}, U]\} \end{cases}$$

Partitioning further on the value of $\mathcal{B} = \frac{K}{Q} - c_m + c_p$ results in the following two subcases:

(a) If $\frac{K}{Q} - c_m + c_p \ge 0$, then Equation (13) can be written as

$$x^*(y) = \begin{cases} -\frac{\mathcal{B}}{\mathcal{A}}, & \xi_1 < y < \xi_2 \\ y, & y \le \xi_1 \text{ or } y \ge \xi_2, \end{cases}$$

for all $y \in [0, U]$. We first show that $\Gamma(-\mathcal{B}/\mathcal{A}, y)$ has a unique maximizer when there are no constraints on y. After much algebra, we can write

$$\Gamma(-\mathcal{B}/\mathcal{A}, y) = \frac{1}{2} \frac{\left(\frac{K}{Q} - c_m + c_p\right)^2}{\left(\frac{r+c_p}{U} - \frac{K}{Q(U-y)}\right)} + \left(r - c_p - \frac{KU}{Q(U-y)}\right)\left(y - \frac{y^2}{2U}\right),$$

whose derivative is

$$\frac{\partial \Gamma(-\mathcal{B}/\mathcal{A}, y)}{\partial y} = \frac{(\frac{K}{Q} - c_m + c_p)^2 U^2 \frac{K}{Q}}{2\left(\frac{KU}{Q} + (U-y)(r+c_p)\right)^2} + \frac{(r-c_p)(U-y)}{U} - \frac{K}{Q} - \frac{\frac{K}{Q} y(2U-y)}{2(U-y)^2}.$$

The first order condition (i.e., $\frac{\partial \Gamma(-\mathcal{B}/\mathcal{A},y)}{\partial y} = 0$) is equivalent to

$$LHS(y) \triangleq \frac{(\frac{K}{Q} - c_m + c_p)^2 U^2 \frac{K}{Q}}{2\left(\frac{KU}{Q} + (U-y)(r+c_p)\right)^2} + \frac{(r-c_p)(U-y)}{U} = \frac{K}{Q} + \frac{\frac{K}{Q} y(2U-y)}{2(U-y)^2} \triangleq RHS(y). \tag{14}$$

Next, since $\frac{\partial LHS(y)}{\partial y} = -\left[\frac{(r+c_p)(\frac{K}{Q}-c_m+c_p)^2 \frac{K}{Q} U^2}{\left((U-y)(c_p+r)+\frac{KU}{Q}\right)^3} + \frac{r-c_p}{U}\right] < 0$ and $\frac{\partial RHS(y)}{\partial y} = \frac{\frac{K}{Q}U^2}{(U-y)^3} > 0$, the $LHS(y)$ function is strictly decreasing in y and the function $RHS(y)$ is strictly increasing in y. If $LHS(0) = \frac{(\frac{K}{Q}-c_m+c_p)^2 \frac{K}{Q}}{2\left(\frac{K}{Q}+(r+c_p)\right)^2} + (r-c_p) < RHS(0) = \frac{K}{Q}$, then Equation (14) has no solution, the derivative is always negative, and the unique maximizer is $y^* = \xi_1$, $x^* = \left(\frac{K}{Q} - c_m + c_p\right) / \left(\frac{r+c_p}{U} - \frac{K}{Q(U-\xi_1)}\right)$, and $\pi_M^2 = \frac{1}{2}\frac{\left(\frac{K}{Q}-c_m+c_p\right)^2}{\left(\frac{r+c_p}{U}-\frac{K}{Q(U-\xi_1)}\right)} + \left(r - c_p - \frac{KU}{Q(U-\xi_1)}\right)\left(\xi_1 - \frac{\xi_1^2}{2U}\right)$. Otherwise, $\Gamma(-\mathcal{B}/\mathcal{A}, y)$ is unimodal with a unique maximum, which is ϕ, the unique root of Equation (14); in other words, ϕ maximizes $\Gamma(-\mathcal{B}/\mathcal{A}, y)$ when there are no constraints on y. Consequently, we conclude that

$$\chi_1 \triangleq \max\{\xi_1, \min\{\phi, \xi_2\}\} = \arg \sup_{\{\xi_1 < y < \xi_2\}} \Gamma(-\mathcal{B}/\mathcal{A}, y).$$

Recalling that $\left(\frac{r-c_m}{2r}\right) U$ maximizes $\Gamma(y,y)$ when there are no constraints on y, we also conclude, similarly to case (1b), that

$$\chi_2 = \arg \max_{\{y \le \xi_1 \text{ or } y \ge \xi_2\}} \Gamma(y,y), \text{ where } \chi_2 \triangleq \begin{cases} \xi_1, & \left(\frac{r-c_m}{2r}\right) U \le \xi_1 \\ \xi_2, & \left(\frac{r-c_m}{2r}\right) U \ge \xi_1 \\ \arg\max_{y \in \{\xi_1, \xi_2\}} \Gamma(y,y). & \end{cases}$$

Finally,

$$y^* = \begin{cases} \chi_1, & -\frac{1}{2}\frac{\left(\frac{K}{Q}-c_m+c_p\right)^2}{\left(\frac{K}{Q(U-\chi_1)}-\frac{(r+c_p)}{U}\right)} - \left(\frac{KU}{Q(U-\chi_1)} - r + c_p\right)\left(\chi_1 - \frac{\chi_1^2}{2U}\right) \geq \left(\frac{r}{U}(U-\chi_2) - c_m\right)\chi_2 \\ \chi_2, & \text{otherwise.} \end{cases}$$

(b) If $\frac{K}{Q} - c_m + c_p < 0$, recall that we showed above that $\xi_1 < 0$, which implies that we can write Equation (13), for $y \in [0, U]$, as

$$x^*(y) = \begin{cases} 0, & \{0 \leq y < \xi_2\} \text{ or } \{y \geq \xi_2 \text{ and } y \in [0, \hat{y}]\} \\ y, & y \in (\max\{\xi_2, \hat{y}\}, U]. \end{cases}$$

We can recursively partition once more, based on the relative values of ξ_2 and $\hat{y}$ (which only depend on problem data):

i. If $\xi_2 \leq \hat{y}$, then

$$x^*(y) = \begin{cases} 0, & y \in [0, \hat{y}] \\ y, & y \in (\hat{y}, U]. \end{cases}$$

In this case,

$$\Gamma(x^*(y), y) = \begin{cases} \left(r - c_p - \frac{KU}{Q(U-y)}\right)\left(y - \frac{y^2}{2U}\right), & y \in [0, \hat{y}] \\ \left(\frac{r}{U}(U-y) - c_m\right) y, & y \in (\hat{y}, U]. \end{cases}$$

We first perform analysis of the two possible forms of the objective, without constraints. The second expression, without the restriction $y \in (\hat{y}, U]$, was analyzed in part (1). We now analyze the first expression without the

restriction $y \in [0, \hat{y}]$. The derivative of this expression can be calculated as

$$\begin{aligned}\frac{\partial \Gamma(0,y)}{\partial y} &= -\frac{KU}{Q(U-y)^2}\left(y - \frac{y^2}{2U}\right) + \frac{1}{U}(U-y)\left(r - c_p - \frac{KU}{Q(U-y)}\right) \\ &= \frac{1}{(U-y)^2}G(y),\end{aligned}$$

where $G(y) \triangleq -\frac{KU}{Q}\left(y - \frac{y^2}{2U}\right) + \frac{1}{U}(r-c_p)(U-y)^3 - \frac{K}{Q}(U-y)^2$. To show that $\Gamma(0,y)$ has a unique maximizer, it suffices to show that $\frac{\partial \Gamma(0,y)}{\partial y} = 0$ has one root, and that the first derivative's sign is positive before the root and negative after the root. To show this, we focus on the $G(y)$ function.

Note that $G(y)$ is cubic in y. Using the standard theory of cubic equations, the discriminant of $G(y)$ is $-\left(\frac{K}{2Q}U\right)^2\left(27(r-c_p)^2 + \frac{K^2}{Q^2}\right) < 0$, which implies that $G(y)$ (and $\frac{\partial \Gamma(0,y)}{\partial y} = 0$) has only one real root, which we denote as ψ:

$$\begin{aligned}\psi \triangleq U - \frac{U}{6}\Bigg(&\frac{K}{Q(r-c_p)} + \frac{\left(\frac{K}{Q}\left(\sqrt{27(r-c_p)^2 + \frac{K^2}{Q^2}} - \sqrt{27}(r-c_p)\right)^2\right)^{\frac{1}{3}}}{r-c_p} \\ &+ \frac{\frac{K^2}{Q^2}}{(r-c_p)\left(\frac{K}{Q}\left(\sqrt{27(r-c_p)^2 + \frac{K^2}{Q^2}} - \sqrt{27}(r-c_p)\right)^2\right)^{\frac{1}{3}}}\Bigg).\end{aligned}$$

Moreover, because $\lim_{y\to-\infty} G(y) = \infty > 0$ and $G(U) = -\frac{KU^2}{2Q} < 0$, then $\psi \in (-\infty, U]$ and ψ maximizes $\Gamma(0,y)$, when there are no constraints on y.

Reincorporating the constraints, we see that

$$\zeta_1 \triangleq \max\{0, \min\{\psi, \hat{y}\}\} = \arg\max_{y\in[0,\hat{y}]}\left(r - c_p - \frac{KU}{Q(U-y)}\right)\left(y - \frac{y^2}{2U}\right), \tag{15}$$

$$\zeta_2 \triangleq \max\left\{\hat{y}, \left(\frac{r-c_m}{2r}\right)U\right\} = \underset{y\in(\hat{y},U]}{\arg\max}\left(\frac{r}{U}(U-y)-c_m\right)y, \qquad (16)$$

and

$$y^* = \begin{cases} \zeta_1, & \left(r-c_p-\frac{KU}{Q(U-\zeta_1)}\right)\left(\zeta_1-\frac{\zeta_1^2}{2U}\right) \geq \left(\frac{r}{U}(U-\zeta_2)-c_m\right)\zeta_2 \\ \zeta_2, & \text{otherwise.} \end{cases}$$

ii. If $\xi_2 > \hat{y}$, then

$$x^*(y) = \begin{cases} 0, & y \in [0, \xi_2) \\ y, & y \in [\xi_2, U]. \end{cases}$$

The analysis is identical to that of case (2bi), except with $\hat{y}$ replaced with ξ_2.

Finally, from Proposition 4.4, the retailer profit is

$$\pi_R^{3D}(w_m^*, w_p^*) = \pi_R^b\left(r-\frac{K}{Q(1-y^*/U)}, r(1-x^*/U)-\frac{K(1-x^*/U)}{Q(1-y^*/U)}\right)+\pi_R^b\left(\frac{K}{Q(1-y^*/U)}, \frac{K}{Q}\right).$$

Note that, for demand uniformly distributed on $[0, U]$, $\pi_R^b(r, w) = \frac{U}{2r}(r-w)^2$. Thus,

$$\begin{aligned} & \pi_R^b\left(r-\frac{K}{Q(1-y^*/U)}, r(1-x^*/U)-\frac{K(1-x^*/U)}{Q(1-y^*/U)}\right) \\ = & \frac{U}{2\left(r-\frac{K}{Q(1-y^*/U)}\right)}\left(r-\frac{K}{Q(1-y^*/U)}-r(1-x^*/U)-\frac{K(1-x^*/U)}{Q(1-y^*/U)}\right)^2 \\ = & \frac{(x^*)^2}{2}\left(\frac{r}{U}-\frac{K/Q}{U-y^*}\right) \end{aligned}$$

and

$$\begin{aligned}\pi_R^b\left(\frac{K}{Q(1-y^*/U)},\frac{K}{Q}\right) &= \frac{U}{2\left(\frac{K}{Q(1-y^*/U)}\right)}\left(\frac{K}{Q(1-y^*/U)}-\frac{K}{Q}\right)^2\\ &= \frac{(y^*)^2K/Q}{2(U-y^*)},\end{aligned}$$

which completes the proof. □ □

Proof. **Lemma 4.3**. Using Propositions 4.3 and 4.4, π_M^2 can be written as

$$\begin{aligned}\max_{w_m,w_p\geq 0}\quad & w_pE[q_p^*(D)]\\ \text{s.t.}\quad & \tfrac{K}{Q}\leq w_m-w_p-c_p\\ & q_p^*(D)=\min\left\{\left(1-\tfrac{K}{(r-w_p-c_p)Q}\right)U,D\right\}\\ & \pi_R^b\left(r-w_p-c_p,\tfrac{K}{Q}\right)\geq 0.\end{aligned}$$

Since w_m does not impact the objective, we set $w_m=\frac{K}{Q}+w_p+c_p$, which eliminates the first constraint. For $(w_m,w_p)\in P_3$, $\frac{K}{Q}\leq w_m-w_p-c_p\Rightarrow\frac{K}{Q}\leq r-w_p-c_p$, which implies that $\pi_R^b\left(r-w_p-c_p,\frac{K}{Q}\right)\geq 0$, and the third constraint is satisfied.

As in the proof of Lemma 4.2, we make the following change of variable: $y=\left(1-\frac{K}{(r-w_p-c_p)Q}\right)U$, with inverse $w_p=r-c_p-\frac{K}{Q(1-y/U)}$. Thus, π_M^2 can be written as

$$\max_{0\leq y\leq U}\left(r-c_p-\frac{K}{Q(1-y/U)}\right)E[\min\{y,D\}]=\max_{0\leq y\leq U}\left(r-c_p-\frac{KU}{Q(U-y)}\right)\left(y-\frac{y^2}{2U}\right),$$

since $E[\min\{y,D\}]=\frac{1}{U}\left(\int_0^y\xi d\xi+\int_y^U(y-x)d\xi\right)=y-\frac{y^2}{2U}$. This problem was analyzed in case (2bi) of the proof of Lemma 4.3. The retailer profit is also calculated in the proof of Lemma 4.2. □ □

Proof. **Proposition 4.6**. We first analyze case 1. If $(r+c_m)^2-4(r+c_p)\frac{K}{Q}<0$, the

maximized profit of case (1) of Lemma 4.2 is identical to that of Lemma 4.1: $\pi_M^2 = \pi_M^1$; thus we must simply compare π_M^1 and π_M^3 to obtain the equilibrium.

We next consider the case $(r + c_m)^2 - 4(r + c_p)\frac{K}{Q} \geq 0$. If $\frac{K}{Q} - c_m + c_p < 0$, note that case (2bi) of Lemma 4.2 returns the maximum of the optimizations defined in Equations (15) and (16). Note that the optimization in Equation (15) has the same objective function as Lemma 4.3, but the feasible region is $y \in [0, \hat{y}]$, rather than $y \in [0, U]$ in Lemma 4.3; since we previously showed that $\hat{y} \leq U$, Lemma 4.3 provides at least as much profit as the first optimization of case (2bi) of Lemma 4.2. Similarly, the optimization in Equation (16) has the same objective function as Lemma 4.1, but the feasible region is $y \in (\hat{y}, U]$, rather than $y \in [0, U]$ in Lemma 4.1; since $\hat{y} \geq 0$, Lemma 4.1 provides at least as much profit as the second optimization of case (2bi) of Lemma 4.2. Consequently, we may ignore case (2bi) of Lemma 4.2. An identical argument, except replacing $\hat{y}$ with $\xi_2 \in [0, U]$, can be made, which eliminates case (2bii) of Lemma 4.2. In summary, we can safely ignore case (2b) of Lemma 4.2. □ □