# ANALYSIS AND DESIGN OF FLEXIBLE SYSTEMS TO MANAGE DEMAND UNCERTAINTY AND SUPPLY DISRUPTIONS

GEOFFREY BRYAN ANG CHUA

NATIONAL UNIVERSITY OF SINGAPORE

2009

# ANALYSIS AND DESIGN OF FLEXIBLE SYSTEMS TO MANAGE DEMAND UNCERTAINTY AND SUPPLY DISRUPTIONS

**GEOFFREY BRYAN ANG CHUA**

(*M.Sci., University of the Philippines*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF DECISION SCIENCES

NATIONAL UNIVERSITY OF SINGAPORE

2009

# ACKNOWLEDGMENT

First of all, I would like to express my sincerest gratitude to my advisor Prof. Mabel Chou. This thesis would not have been possible without her continuous support and guidance. I am fortunate to know Prof. Chung-Piaw Teo as my mentor, and I thank him for sharing with me his knowledge and passion for research. It is a great honor for me to have spent the past five years learning from them.

I am thankful to my thesis committee members, Prof. Melvyn Sim and Prof. Sun Jie, for their valuable suggestions and guidance throughout my Ph.D. study. Profs. James Ang, Rick So, Chou Fee Seng, Yaozhong Wu, Jihong Ou, and Hengqing Ye at the Decision Sciences department, and Profs. Andrew Lim, George Shanthikumar, and Max Shen at Berkeley have also taught me many things about research and academic life in general.

I am specially grateful to Marilyn Uy and Victor Jose, two long-time friends with whom I shared the same academic path for the past five years. It was our friendship and mutual encouragement that got me through some tough times. Our friendship is truly a blessing. I also want to thank my friends at NUS, Huan Zheng, Wenqing Chen, Hua Tao, Shirish Srivastava, Annapoornima Subramaniam, Marcus Ang, Su Zhang, Qingxia Kong, Vinit Kumar, and Zaheed Halim, for the exciting times and wonderful memories.

I will forever be indebted to my parents for their nurture and unconditional love. Likewise, I am thankful to my siblings Irene, Stanley, Catherine and Frederick for their support and encouragement.

Finally, I express my heartfelt gratitude, love and admiration to my fiancée Gem, whose love and support have been a source of joy and a pillar of strength for me.

G. A. Chua

Singapore, April 2009

# CONTENTS

# ABSTRACT

Facing intense market competition and high demand variability, firms are beginning to use flexible process structures to improve their ability to match supply with uncertain demand. The concept of chaining has been extremely influential in this area, with many large automakers already making this the cornerstone of their business strategies to remain competitive in the industry. In this thesis, we aim to provide a theoretical justification for why partial flexibility works nearly as well as full flexibility. We also seek to extend the theory of partial flexibility to environments that take into account new factors relevant to the practice of process flexibility.

We first study the asymptotic performance of the chaining strategy in the symmetric system where supply and (mean) demand are balanced and identical. We utilize the concept of a generalized random walk to show that an exact analytical method exists that obtains the chaining efficiency for general demand distributions. For uniform and normal demand distributions, the results show that the 2-chain already accrues at least 58% and 70%, respectively, of the benefits of full flexibility. Our method can also be extended to more general cases such as non-symmetrical demands, unbalanced systems, and higher-degree chains.

We then extend our analysis to take into account the response dimension,

the ease with which a flexible system can switch from producing one product to another. Our results show that the performance of any flexible system may be seriously compromised when response is low. Nevertheless, our analytical lower bounds show that under all response scenarios, the 2-chain still manages to accrue non-negligible benefits (at least 29.29%) vis-à-vis full flexibility. Furthermore, we find that given limited resources, upgrading system response outperforms upgrading system range in most cases, suggesting a proper way to allocate resources. We also observe that improving system response can provide even more benefits when coupled with initiatives to reduce demand variability.

Next, we consider the impact of partial production postponement on the performance of flexible systems. Under partial postponement, we find that results on chaining under full postponement may not hold. In the example of small systems, when postponement level is lower than 80%, the celebrated 2-chain may perform quite badly, with a performance loss of more than 12%. By adding another layer of flexibility, i.e. a third chain, the optimality loss is restored to 5% even when postponement drops to 65%. We also study the flexibility-postponement tradeoff and find that a firm operating with a 3-chain at 70% postponement can perform extremely well with minimal optimality loss.

Finally, we look into the fragility of flexible systems under the threat of supply disruptions. Under both link and node disruptions, we find that having a third chain, or a third layer of flexibility in the asymmetric setting, can greatly reduce system fragility. Furthermore, when additional capacity is made available, the performance of the third chain appears to be insensitive

to how this extra capacity is allocated, which differs from the case of the 2-chain. These observations, in conjunction with the recommendations for partial production postponement, suggest that there is substantial value in employing the third chain.

# LIST OF FIGURES

# LIST OF TABLES

# 1. INTRODUCTION

Since the 1980s, we have witnessed the advent of globalization and the tremendous effects it has on world consumption and production. A quick look at a BusinessWeek report [2] on the top 100 brands in 2007 reveals that these brands already hail from twelve different countries around the world. (See Table 1.1 for a partial listing.) According to the report, each of these brands derives at least a third of its earnings outside its home country. This tells us that increasingly, the world is moving towards a phenomenon of borderless consumption. That is, for consumers, the world is becoming their shopping mall. On the other hand, for manufacturers, the whole world is becoming their customer.

With the said internationalization of market competition, firms nowadays need to build up the capacity for becoming competitive as a world-class company. The most common solution has been to turn to outsourcing and offshoring, essentially tapping into the production capabilities of factories, big and small, all over the world. For example, many American and European brands outsource their sourcing function to Hong Kong-based Li & Fung, the world's leading supply chain company who controls a network of over 10,000 production facilities scattered everywhere in places like China, Brazil, the Czech Republic, Honduras, Mauritius, Mexico, Poland, South Africa,

| Country | Brand(s) |
|---|---|
| United States | Coca-Cola, Microsoft, Nike, Disney, Apple, Starbucks |
| Japan | Toyota, Canon, Nintendo, Sony |
| Finland | Nokia |
| Germany | BMW, Siemens, SAP, Adidas, Nivea |
| France | Louis Vuitton, AXA, L'Oreal, Hennessy, Chanel |
| South Korea | Samsung, Hyundai, LG |
| Britain | HSBC, Reuters, BP, Smirnoff, Burberry |
| Switzerland | Nescafe, UBS, Nestle, Rolex |
| Sweden | IKEA |
| Netherlands | Philips, ING |
| Italy | Gucci, Prada |
| Spain | Zara |

*Tab. 1.1:* Partial Listing of Top 100 Brands by Country

Zimbabwe, and countries in Southeast Asia [21]. On this phenomenon of borderless manufacturing, Fung et al [24], [25] believe the trend is "to rip the roof off the factory. In contrast to Henry Ford's assembly line, where all the manufacturing processes were under one roof, the entire world is our factory." Other than granting firms the ability to increase capacity through global aggregation, this strategy also allows the firms to control and reduce operating expenses as well as focus on improving their core businesses, such as product design and marketing.

Another important trend is the fragmentation of consumer demand. Instead of catering to one big market with more or less homogeneous demand, companies are beginning to see more niche markets with diverse tastes as well as the emergence of variety-seeking consumer behavior. As this trend becomes more prevalent, we see an increasing proliferation of product lines as companies struggle to stay competitive. In the automobile industry, the number of car models offered in the United States market has increased from

195 (in 1984), to 238 (in 1994), to 282 (in 2004), and was projected to reach 330 by 2008 (cf. [54]). The same phenomenon can be observed in other industries such as electronics, clothing, food products, and even services like entertainment/media and education. As a result, demand uncertainty on a per product basis increases and forecasting becomes more difficult.

Facing such an increased demand uncertainty as well as heightened market competition, businesses can no longer rely on capacity, pricing, quality, and timeliness alone as competitive strategies. One approach in recent years that has proven effective is the use of flexible production facilities. In the automobile industry, for example, companies are increasingly moving from focused factories to flexible factories. According to a survey conducted in 2004, the plants of major automobile manufacturers in North America, such as Ford and General Motors, are more flexible than their counterparts 20 years ago (cf. [53]). The survey shows that these flexible plants can produce many more types of cars to cater to rapidly changing consumer demands while the plant capacities have not changed very much. The kind of flexibility adopted in these plants is known as "process flexibility" in the operations management literature.

## 1.1 Process Flexibility

"Process flexibility" can be defined as a firm's ability to provide varying goods or services, using different facilities or resources (cf. [32], [47]). Nowadays, it has become a common strategy among players in the automobile industry to employ process flexibility in their production facilities [53]. This focus on

process flexibility as a competitive strategy can likewise be observed in other manufacturing industries, such as the textile/apparel industry [19] and the semiconductor/electronics industry [43]. The value of flexibility also extends to service industries, where firms have increasingly employed cross-trained workers to provide more flexible services [30].

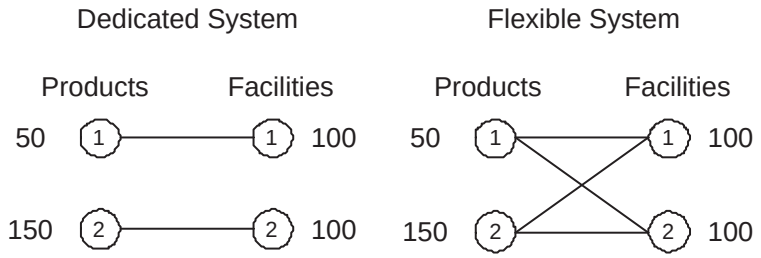


*Fig. 1.1:* The Benefits of Process Flexibility

To illustrate the benefits gained from employing process flexibility, we must first understand how a flexible production system works. Consider the two systems in Figure 1.1. Both systems have two products and two facilities. The demands of the products are random while the capacities of the facilities are fixed at 100 units each. The system on the left is a dedicated production system (also known as a focused factory) while the one on the right is a flexible system. When demand for product 1 is low while demand for product 2 is high, the extra demand for product 2 is lost to the dedicated system and the extra capacity of facility 1 is wasted. On the other hand, a flexible system is able to recover an additional sales of 50 units due to its ability to produce more products in each facility. This is the fundamental reason why process flexibility has been an effective strategy in many industries. In an interview with the Wall Street Journal [11], Chrysler Group CEO Thomas LaSorda

disclosed that flexible production "gives us a wider margin of error." With regard to the value of process flexibility, he said, "if the Caliber doesn't sell well, the Jeep Compass and Patriot could take up capacity, and eventually a fourth model will be built, too."

The theoretical justification for the effectiveness of process flexibility can be traced back to the early work of Eppen [20]. For a multi-location newsvendor problem, he showed that the mismatch cost for a decentralized system exceed those in a centralized system, and that the gap between these two systems depends on the demand correlation. Indeed, a decentralized system is analogous to a dedicated production system, while the centralized system corresponds to flexible production. Likewise, it makes sense that process flexibility is most effective when product demands are negatively correlated and least effective when demand correlation is positive.

It should be noted, however, that Eppen's result on the benefits of consolidation or risk pooling is predicated on the assumption of full consolidation or complete pooling. In the context of process flexibility, we must have a fully flexible production system where all facilities can produce all products for the said theory to hold. In addition, most of the early works on process flexibility examine the appropriate mix of dedicated versus flexible resources, thus focusing only on fully flexible resources. Unfortunately, many companies realize that full flexibility typically comes at great expense, thus they can only make limited use of these theories on full flexibility. This calls for a new or extended theory of partial flexibility.

With most facilities capable of producing most products, one may overinvest in process flexibility. On the other hand, when one has too little or

no flexibility at all, this may result in a high level of lost sales. This becomes a question of striking a balance between flexibility and cost, which can be restated as whether one can achieve the benefits of full flexibility at an acceptable cost level. Jordan and Graves [32] show via simulation studies that this is possible using the concept of a simple "chaining" strategy. Here, a plant capable of producing a small number of products, but with proper choice of the **process structure** (i.e., plant-product linkages), can achieve nearly as much benefit as the full flexibility system. This concept is widely believed to be true, and has been applied successfully in many industries. For example, Chrysler CEO LaSorda has repeatedly mentioned the importance of chaining in his interviews and speeches [35], while VP Frank Ewasyshyn was recently inducted into the Shingo Prize Academy for his contributions to flexibility and efficiency [1]. Jordan and Graves [32] also applied the chaining strategy to General Motors' production network.

To enhance our understanding of the progress in this research and to put in perspective the contributions of this thesis, a thorough literature review on process flexibility is provided in Section 1.1.1.

### *1.1.1 Literature Review*

In the operations management literature, there are two main streams of research related to process flexibility. The first stream examines the trade-off between flexible and dedicated resources. Fine and Freund [22] characterize the optimal investment in flexibility (i.e. the optimal amounts of dedicated and flexible resources) for a price-setting firm, where demand is modeled by

a discrete probability distribution of $k$ possible states that affect demand. Van Mieghem [55] takes a critical-fractile approach to solving the optimal flexibility investment for a price-taking firm, but for any arbitrary multivariate demand distribution. Bish and Wang [10] extend van Mieghem's work to a price-setting firm facing different types of correlated demands.

The above studies, though, focus only on full flexibility; that is, all facilities can produce all types of products. Unfortunately, in practice, the acquisition cost of full flexibility is usually too enormous to permit the recovery of adequate benefits. In response, a second stream of research looks at different degrees of flexibility, and examines the value of these types of process flexibility. The landmark study was by Jordan and Graves [32], who introduced the concepts of "smart limited flexibility" and "chaining". They observe, through extensive simulation, that limited flexibility, configured the right way, yields most of the benefits of full flexibility. Furthermore, they claim that limited flexibility has the greatest benefits when a "chaining" strategy is used. In the **symmetric case** where the (mean) demand and facility capacity are balanced and identical, a chaining configuration is formed by enabling every facility to produce two products and every product to be produced by two facilities, in a way that "chains" up all the facilities and products. For a 10-facility, 10-product example, the expected sales generated from chaining is compared to that of full flexibility using numerical simulation. The results show that chaining already achieves about 95% of the benefits of full flexibility while incurring only a small fraction of the cost. Figure 1.2 provides an illustration.

The theory developed and the insights gained from studying the sym-

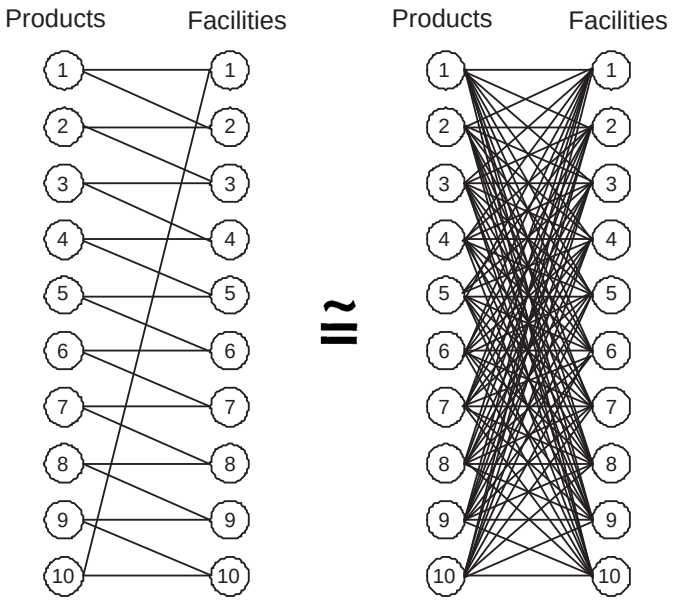


*Fig. 1.2:* Chaining is Almost as Good as Full Flexibility

metric case are then used to formulate principles and guidelines to address the more sophisticated **asymmetric case** where facilities can have varying capacities while product demands may follow arbitrary probability distributions. Here, Jordan and Graves follow similar ideas of adding more linkages to the system such that the resulting structure forms a cycle (albeit not necessarily a regular chain). In addition, they propose a probabilistic measure (later called the JG index) that can be used for evaluating different flexibility structures. Applying these concepts to General Motors' production network, they find that indeed a partially flexible system, if well designed, already captures almost all the benefits of full flexibility.

Because the twin ideas of smart limited flexibility and chaining have been well received, many researchers subsequently applied and examined these strategies in various other contexts such as supply chains ([27], [10]),

queuing ([7], [28]), revenue management ([26]), transshipment distribution network design ([39], [58]), manufacturing planning ([34]) and flexible work force scheduling ([18], [30], [57], [13]). For example, Graves and Tomlin [27] extended the study to multi-stage supply chains and found that "chaining" also works very well. Hopp et al. [30] observed similar results in their study of a work force scheduling problem in a ConWIP (constant work-in-process) queuing system. They compared "cherry picking", where capacity is "picked" from all other stations versus "skill-chaining" where workforce in each station is cross-trained to perform work in the next adjacent station. They observed that "skill-chaining" outperforms "cherry picking" and also that a chain with a low degree (the number of tasks a worker can handle) is able to capture the bulk of the benefits of a chain with high degree.

Another issue addressed in the literature is the search for effective indices to measure the performance of flexibility structures (cf. [32], [27], [31], and [17]). For example, Jordan and Graves [32] proposed a probabilistic index, which roughly measures the probability that unsatisfied demand from a subset of products in a given flexible system would exceed that of a fully flexible system. However, this index is usually very hard to compute if demands are not normally distributed or they are correlated due to the complexity of the joint probability distribution. This renders the index of limited use especially in the case of correlated demands when such performance indices are most needed. To overcome this problem, Iravani et al. [31] proposed a new perspective on flexibility using the concept of "structural flexibility" and introduced new flexibility indices. The indices are obtained by first defining the "structural flexibility matrix" and then taking the largest eigenvalue as

well as the mean of this matrix as flexibility indices. These indices are easy to compute and are applicable to serial, parallel, open, and closed networks. More recently, Chou et al. [17] introduced the Expansion Index, based on the concept of graph expander. They define this index as the second smallest eigenvalue of an associated Laplacian matrix. Numerical experiments show that this index performs as well, if not better than the previous indices in most of the problem instances considered.

Another group of studies tries to warn the community about some unaccounted issues when employing process flexibility. Bish et al. [9] go beyond just matching supply and demand as they study the impact of flexibility on the supply chain. They show that in a $2 \times 2$ system, certain practices that may seem reasonable in a flexible system can result in greater production swings and higher component inventory levels, which will then increase operational costs and reduce profits. To account for partial flexibility, Muriel et al. [45] extend Bish et al.'s work to larger systems and obtain similar findings. Brusco and Johns [13] present an integer linear programming model to evaluate different cross-training configurations in a workforce staffing problem. In their model, they consider a case wherein a worker is 100% efficient in his primary skill but only 50% efficient in his secondary skill. Under this scenario, the value of skill-chaining may be significantly reduced due to the efficiency lost in using secondary capacity. In this thesis, we also examine issues and concerns not previously considered in the literature. At the same time, we propose measures on how to mitigate the effects of these additional factors. We defer this discussion to Section 1.2.

The previous works cited above present limited concrete analytical re-

sults. To strengthen the analytical aspect, Akşin and Karaesmen [3] first show that the optimal system sales for any demand realization in a given flexible system can be obtained by deriving the maximum flow in a network flow model. The performance of the system (in terms of expected sales) is therefore equivalent to determining the expected amount of maximum flow in a network with random capacities. The authors then use their network flow model to show that the expected throughput is concave in the degree of flexibility. This implies the diminishing value of additional flexibility, partly explaining why chaining already gives a substantial portion of the benefits of full flexibility. Bassamboo et al. [6] study the optimal type and amount of flexibility for stochastic processing systems. Focusing on high-volume symmetric systems and using heavy-traffic queueing analysis, they analytically demonstrate that the optimal flexibility configuration invests a lot in dedicated resources, a little in only bi-level flexibility, but nothing in level-k $> 2$ flexibility, let alone full flexibility. Chou et al. [17] use the concept of graph expanders to provide a rigorous proof of the existence of a sparse partially flexible structure (not necessarily chaining) for a symmetrical system that accrues most of the benefits of full flexibility. In another paper, Chou et al. [16] use constraint sampling to characterize the analytical performance of sparse structures, vis-à-vis the full flexibility system, when the demand and supply are asymmetrical. However, no theoretical results exist on how to analytically capture exactly how well the chaining strategy performs.

As mentioned, the process flexibility problem is intimately related to the problem of determining the expected amount of maximum flow in a network with random capacity. Karp et al. [33] developed an algorithm to find

the maximum flow in a random network with high probability, but to the best of our knowledge, the algorithm could not be used to find the expected maximum flow value. For the case when the capacities are exponentially distributed, Lyons et al. [41] used the connection between random walk and electrical network theory to bound the expected max flow value by the conductance of a related electrical network (where the capacity of each arc is replaced by the expected capacity value). The proof technique relies heavily on the properties of the exponential distribution and thus cannot be utilized for more general distribution. Hence, a non-simulation-based method for obtaining the expected maximum flow in the random network of process flexibility must be developed from scratch.

## 1.2 Research Objectives and Results

The objectives of this thesis are:

- *To provide further theoretical justification for the effectiveness of the chaining strategy:* Although some works have already started toward building the theory of partial flexibility, it remains to be established exactly how effective the chaining strategy is. The classical simulation result by Jordan and Graves that a 2-chain in a $10 \times 10$ system already captures 95% of the benefits of full flexibility has yet to be justified or reproduced analytically. We utilize the concept of a generalized random walk to show that an exact analytical method exists that obtains the chaining efficiency for very large systems. This method works for a wide range of demand distributions and confirms the belief in the community

that chaining is almost as good as full flexibility. More importantly, our proposed method can be generalized and incorporated into the analysis of more sophisticated settings.

- *To examine the performance of chaining as system size grows infinitely large:* For small $n$ (say $n = 10$), previous works already show that chaining accrues about 95% of the benefits of full flexibility. As system size increases, this value tends to decrease based on our additional simulations. A natural question would then be how fast chaining performance deteriorates as $n$ increases to infinity. Such asymptotic analysis is important given today's growing manufacturing and service networks, and complements existing literature which is largely simulation based and thus confined only to small or moderate size systems. Our proposed random walk method can be used to obtain exact analytical values for the asymptotic chaining efficiency. These values also serve as lower bounds for any finite system size $n$. Interestingly, even when system size is infinitely large, our results show that chaining can still offer most ($70\% \gg 0$) of the benefits of full flexibility.

- *To examine the performance of chaining when system response is not perfect:* It has been suggested in the literature and confirmed among managers that process flexibility must be viewed based on two dimensions: range and response. Range is the set of states that a system can adopt, while response is the ease with which the system switches from state to state. Although both dimensions are important, the existing literature does not analytically examine the response dimension – most

works assume system response is always perfect. We model the response dimension in terms of production efficiencies such that primary production is less expensive (more efficient) than secondary production. We use the Max-Flow Min-Cut theorem to obtain lower bounds in our quest to characterize the chaining performance for all relevant response levels. We can show that the performance of any flexible system may be significantly lowered when operating under low response levels. Nevertheless, our lower bounds show that under all response scenarios, chaining still manages to accrue non-negligible benefits (at least 29.29%) vis-à-vis full flexibility.

- *To examine the performance of chaining under partial production postponement:* Aside from process flexibility, another approach that can help deal with demand uncertainty is production postponement. Production postponement is "the firm's ability to set production quantities after demand uncertainty is resolved". When there is no postponement, the firm acts as a make-to-stock manufacturer; with full postponement, it behaves in a make-to-order fashion. Because existing literature on process flexibility assumes full postponement, we seek to understand how the existing theories hold under partial postponement. We utilize a multi-item newsvendor model with second supply and partial capacity sharing to study both partial flexibility and partial postponement. We find that results on chaining under full postponement may not hold under partial postponement. For small systems, when postponement level is lower than 80%, the celebrated 2-chain may perform quite badly,

with a performance loss of more than 12%. By adding another layer of flexibility, i.e. a third chain, the optimality loss is restored to 5% even when postponement drops to 65%. This serves as evidence for the potential value of employing a third chain (or in the asymmetric case, a third layer of flexibility).

- *To examine the performance of chaining under supply disruptions:* Recent studies have pointed out that supply chains are increasingly susceptible to disruptions that may be caused by labor strikes, hurricanes, fires, and other unexpected calamities. It has been shown that measures used to protect against demand uncertainty and yield uncertainty are not suitable for mitigating disruption risks. Instead, one must equip his supply chains with more redundancy or slack to buffer against disruption uncertainty. However, firms have historically been disinclined to invest in additional infrastructure or inventory, despite the potentially large payoff in the event of a disruption. Hence, it is but natural to turn to process flexibility for a way to reduce the buffer requirements or to maximize the utilization of additional resources. We study the fragility of flexible systems and how it changes when more flexibility is introduced or when additional capacity is provided. We find that the third chain, or a third layer of flexibility in the asymmetric case, can greatly reduce system fragility. It can also increase implementation flexibility in terms of how additional capacity must be allocated.

## 1.3 Preliminaries: Models and Measures

As in existing literature, there are usually two cases considered for the study of partial flexibility: the symmetric case and the asymmetric case. In this thesis, we focus on the symmetric case for the purpose of theory-building. Insights gained from this exercise are then transferred and numerically tested on the asymmetric case. For that reason, we define the general notations for our analysis based on the symmetric setting.

Any flexibility structure for an $n$-product, $n$-facility system can be represented by a bipartite graph $G(n) = (\mathcal{A}(n) \cup \mathcal{B}(n), \mathcal{G}(n))$. On the left is a set $\mathcal{A}(n)$ of $n$ product nodes while on the right is a set $\mathcal{B}(n)$ of $n$ facility nodes. An edge $e = (i, j) \in \mathcal{G}(n)$ connecting product node $i$ to facility node $j$ means that facility $j$ is endowed with the capability to produce product $i$. Here, $\mathcal{G}(n) \subseteq \mathcal{A}(n) \times \mathcal{B}(n)$ denotes the set of all such links; that is, the edge set of the bipartite graph. Hence, each flexibility configuration can be uniquely represented by the edge set $\mathcal{G}(n)$. The three most common flexibility configurations studied in the literature are:

1. The dedicated system:

   $$\mathcal{D}(n) = \{(i, i) \mid i \in \{1, 2, \ldots, n\}\}$$

2. The chaining system[1]:

   $$\mathcal{C}(n) = \{(i, i) \mid i = 1, 2, \ldots, n\} \cup \{(1, 2), (2, 3), \ldots, (n-1, n), (n, 1)\}$$

[1] This structure is also known as the 2-chain (because each plant is connected to two products, and vice versa) or the long chain (because it is the longest possible 2-chain).

3. The full flexibility system:

$$\mathcal{F}(n) = A(n) \times B(n)$$

Figure 1.3 shows some examples of flexibility configurations for a three-facility, three-product system. Graphs (a), (b), and (c) are the three respective special configurations as listed above for the case $n = 3$.

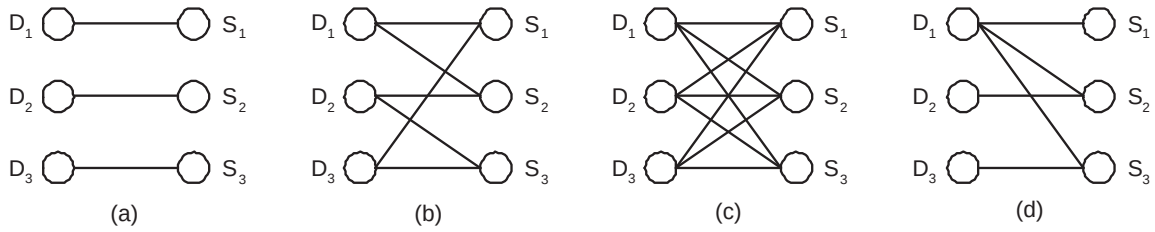


*Fig. 1.3:* Bipartite Graph Representation of $3 \times 3$ Flexibility Structures

We also generalize the above chaining system $\mathcal{C}(n)$ to higher-degree chains. Previously, the degree of each product or facility node in the chaining strategy is set at 2. In general, we can extend this to degree $d \leq n$, where each product node is connected to $d$ facility nodes and each facility node is linked to $d$ product nodes. Clearly, when $d = 1$ and $d = n$, we recover the dedicated and full flexibility systems, respectively. The expanded notation is as follows. For $d = 1, 2, \ldots, n$, the $d$-chain is

$$\begin{aligned}\mathcal{C}_d(n) &= \left\{ \bigcup_{i=1}^{n-d+1} \{(i,i),(i,i+1),\ldots,(i,i+d-1)\} \right\} \\ &\bigcup \left\{ \bigcup_{i=n-d+2}^{n} \{(i,i),(i,i+1),\ldots,(i,n),(i,1),(i,2),\ldots,(i,i-n+d-1)\} \right\}\end{aligned}$$

We use the notation $\mathcal{C}_d(n)$ when comparing the performance of the 2-chain with higher-degree chains. Otherwise, we revert to the original notations $\mathcal{D}(n) = \mathcal{C}_1(n)$, $\mathcal{C}(n) = \mathcal{C}_2(n)$ and $\mathcal{F}(n) = \mathcal{C}_n(n)$.

We let $\mathbf{D} = (D_1, D_2, \ldots, D_n)$ denote the demand vector and $\mathbf{C} = (C_1, C_2, \ldots, C_n)$ denote the supply vector. Each demand $D_i$ is assumed to be random and follow some distribution function $F_i$, while every supply capacity $C_j$ is fixed. In the symmetric case, we further assume that $D_1, D_2, \ldots, D_n$ are i.i.d. and follow the same distribution $F$, whereas all facilities have the same capacity $C_j = C$.

### *1.3.1 Optimization Models*

The problem boils down to solving an optimization model for each realization of product demands $\mathbf{D}$. The expectation of the optimal objective value (whether sales, profit, or cost) is computed and incorporated into performance measures for flexibility structures. The optimization models we consider in this thesis are: (1) the Maximum Flow Model, (2) the Maximum Profit Model, and (3) the Minimum Mismatch Cost Model.

1. *The maximum flow model:* In this model, we find the maximum sales possible given the demand realizations, facility capacities and the flexibility configuration. This is a suitable model when system response is perfect and products have equal unit revenues and unit production costs.

$$
\begin{aligned}
Z^*_{\mathcal{G}(n)}(\mathbf{D}) = \quad & \max \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij} \qquad (1.1)\\
s.t. \quad & \sum_{j=1}^{n} x_{ij} \le D_i \ \forall i = 1, 2, \ldots, n;\\
& \sum_{i=1}^{n} x_{ij} \le C_j \ \forall j = 1, 2, \ldots, n;\\
& x_{ij} \ge 0 \ \forall \ i, j = 1, \ldots, n,\\
& x_{ij} = 0 \ \ \forall \ (i,j) \notin \mathcal{G}(n).
\end{aligned}
$$

2. *The maximum profit model:* This is the model we use to study the response dimension. Because we model system response in terms of production efficiencies, we must consider a maximum profit criterion to account for the more expensive secondary or backup production. Here, we let $p$ be the unit revenue, $c_p$ be the unit cost of primary production, and $c_s (\ge c_p)$ be the unit cost of secondary production.

$$
\begin{aligned}
\Pi^*_{\mathcal{G}(n)}(\mathbf{D}, c_s) = \quad & \max \quad (p - c_p)\sum_{i=1}^{n} x_{ii} + (p - c_s)\sum_{i=1}^{n}\sum_{j\neq i} x_{ij} \quad (1.2)\\
& \text{s.t.} \quad \sum_{j=1}^{n} x_{ij} \le D_i \quad \forall \ i = 1, \ldots, n\\
& \qquad \sum_{i=1}^{n} x_{ij} \le C_j \quad \forall \ j = 1, \ldots, n\\
& \qquad x_{ij} \ge 0 \quad \forall \ i, j = 1, \ldots, n\\
& \qquad x_{ij} = 0 \quad \forall \ (i,j) \notin \mathcal{G}(n)
\end{aligned}
$$

3. *The minimum mismatch cost model:* We use the following multi-item newsvendor model with secondary supply and partial capacity sharing to examine process flexibility under partial production postponement. We let $\alpha$ denote the level of postponement[2]. Hence, the problem becomes a two-stage optimization model where $(1-\alpha)$ of the capacity must be allocated before actual demand is observed while that of the remaining $\alpha$ of the capacity can be postponed after demand is made known. Here, the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ denote first-stage production and second-stage production, respectively. The vector $\boldsymbol{\xi}$ denotes the realization of the demand vector, while $c_o$ and $c_u$ represent the unit overage and underage costs.

$$
\begin{aligned}
G^*_{\mathcal{G}(n)}(\alpha) = \quad & \min_{\boldsymbol{x}} \quad G_{\mathcal{G}(n)}(\boldsymbol{x}, \alpha) \\
& \text{s.t.} \quad \sum_{i=1}^{n} x_{ij} \leq (1-\alpha) C_j \quad \forall j = 1, 2, \ldots n \\
& \qquad\quad x_{ij} \geq 0 \quad \forall i, j = 1, 2, \ldots n \\
& \qquad\quad x_{ij} = 0 \quad \forall (i,j) \notin \mathcal{G}(n)
\end{aligned}
\tag{1.3}
$$

where

$$
G_{\mathcal{G}(n)}(\boldsymbol{x}, \alpha) = c_o \cdot g_1(\boldsymbol{x}) + c_u \cdot g_2(\boldsymbol{x}) - c_u \cdot \mathrm{E}[h_{\mathcal{G}(n)}(\boldsymbol{x}, \alpha, \boldsymbol{\xi})]
$$

$$
g_1(\boldsymbol{x}) = \sum_{i=1}^{n} \int_0^{\sum_{j=1}^{n} x_{ij}} \left( \sum_{j=1}^{n} x_{ij} - \xi_i \right) dF_i(\xi_i)
$$

[2] In practice, $\alpha$ can be different for different products. However, we use the same parameter $\alpha$ for all products in order to have analytical tractability so as to gain insights into the general effect of the postponement level on system performance.

$$g_2(\boldsymbol{x}) = \sum_{i=1}^{n} \int_{\sum_{j=1}^{n} x_{ij}}^{\infty} \left( \xi_i - \sum_{j=1}^{n} x_{ij} \right) dF_i(\xi_i)$$

and

$$\begin{aligned} h_{\mathcal{G}(n)}(\boldsymbol{x}, \alpha, \boldsymbol{\xi}) = \quad & \max_{\boldsymbol{y}} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} y_{ij} \\ & \text{s.t.} \quad \sum_{j=1}^{n} y_{ij} \leq \left( \xi_i - \sum_{j=1}^{n} x_{ij} \right)^{+} \quad \forall i = 1, 2, \ldots n \\ & \qquad \sum_{i=1}^{n} y_{ij} \leq \alpha C_j \quad \forall j = 1, 2, \ldots n \\ & \qquad y_{ij} \geq 0 \quad \forall i, j = 1, 2, \ldots n \\ & \qquad y_{ij} = 0 \quad \forall (i,j) \notin \mathcal{G}(n) \end{aligned}$$

### 1.3.2 Performance Measures

The above optimization problems have to be solved for each realization of demand and the expectation of the optimal objective function values is taken with respect to the demand uncertainty. In fact, this poses as one of the main challenges in our analysis. Nevertheless, once these expected values are obtained, they can be subsequently included in the computation of the following performance measures for different flexibility structures.

1. Expected Sales Ratio: This measures the performance of any partially flexible system in terms of expected sales relative to full flexibility.

$$SR_{\mathbb{P}}(\mathcal{G}(n)) = \frac{\mathrm{E}_{\mathbb{P}}[Z^*_{\mathcal{G}(n)}(\mathbf{D})]}{\mathrm{E}_{\mathbb{P}}[Z^*_{\mathcal{F}(n)}(\mathbf{D})]}$$

where $\mathbb{P}$ is the probability measure that characterizes random demand vector $\mathbf{D}$. Since distributional ambiguity is not the focus of this study, the probability measure in question is usually distinct and clear from the context. Hence, from here onwards, we drop $\mathbb{P}$ for notational simplicity. The simplified notation for the Expected Sales Ratio becomes

$$SR(\mathcal{G}(n)) = \frac{\mathrm{E}[Z^*_{\mathcal{G}(n)}(\mathbf{D})]}{\mathrm{E}[Z^*_{\mathcal{F}(n)}(\mathbf{D})]}$$

2. Expected Benefits Ratio (or Flexibility Efficiency): This measures the performance of any partially flexible system in terms of expected improvements (which may be in terms of sales, profit, or cost) over the dedicated system, relative to full flexibility.

$$FE(\mathcal{G}(n)) = \frac{\mathrm{E}[Z^*_{\mathcal{G}(n)}(\mathbf{D})] - \mathrm{E}[Z^*_{\mathcal{D}(n)}(\mathbf{D})]}{\mathrm{E}[Z^*_{\mathcal{F}(n)}(\mathbf{D})] - \mathrm{E}[Z^*_{\mathcal{D}(n)}(\mathbf{D})]}$$

or

$$FE(\mathcal{G}(n), c_s) = \frac{\mathrm{E}[\Pi^*_{\mathcal{G}(n)}(\mathbf{D}, c_s)] - \mathrm{E}[\Pi^*_{\mathcal{D}(n)}(\mathbf{D}, c_s)]}{\mathrm{E}[\Pi^*_{\mathcal{F}(n)}(\mathbf{D}, c_s)] - \mathrm{E}[\Pi^*_{\mathcal{D}(n)}(\mathbf{D}, c_s)]}$$

or

$$FE(\mathcal{G}(n), \alpha) = \frac{G^*_{\mathcal{D}(n)}(\alpha) - G^*_{\mathcal{G}(n)}(\alpha)}{G^*_{\mathcal{D}(n)}(\alpha) - G^*_{\mathcal{F}(n)}(\alpha)}$$

whichever is appropriate in the given context.

3. Chaining Efficiency: This is a shorthand for the flexibility efficiency of

the chaining system or the 2-chain.

$$CE(n) = FE(\mathcal{C}(n))$$

or

$$CE(n, c_s) = FE(\mathcal{C}(n), c_s)$$

or

$$CE(n\alpha) = FE(\mathcal{C}(n), \alpha)$$

whichever is appropriate in the given context.

4. Chaining Efficiency for d-chains: This is a shorthand for the flexibility efficiency of a $d$-chain.

$$CE_d(n) = FE(\mathcal{C}_d(n))$$

or

$$CE_d(n, c_s) = FE(\mathcal{C}_d(n), c_s)$$

or

$$CE_d(n, \alpha) = FE(\mathcal{C}_d(n), \alpha)$$

whichever is appropriate in the given context.

5. Asymptotic Sales Ratio: This is the asymptotic limit of the Expected

Sales Ratio as system size expands to infinity.

$$ASR(\mathcal{G}(\infty)) = \lim_{n\to\infty} SR(\mathcal{G}(n))$$

6. Asymptotic Chaining Efficiency: This is the asymptotic limit of the Chaining Efficiency as system size expands to infinity.

$$ACE = \lim_{n\to\infty} CE(n)$$

or

$$ACE(c_s) = \lim_{n\to\infty} CE(n, c_s)$$

or

$$ACE(\alpha) = \lim_{n\to\infty} CE(n, \alpha)$$

whichever is appropriate in the given context.

7. Asymptotic Chaining Efficiency for d-chains: This is the asymptotic limit of the Chaining Efficiency of a $d$-chain as system size expands to infinity.

$$ACE_d = \lim_{n\to\infty} CE_d(n)$$

or

$$ACE_d(c_s) = \lim_{n\to\infty} CE_d(n, c_s)$$

or

$$ACE_d(\alpha) = \lim_{n\to\infty} CE_d(n, \alpha)$$

whichever is appropriate in the given context.

8. Optimality Loss/Gap: This measures the loss in a system with partial flexibility and partial postponement relative to the optimal system which possesses full flexibility and full postponement.

$$OG(\mathcal{G}(n), \alpha) = \frac{G^*_{\mathcal{G}(n)}(\alpha)}{G^*_{\mathcal{F}(n)}(1)} - 1$$

## 1.4 Structure of Thesis

The remaining sections of the thesis are organized as follows. The asymptotic chaining efficiency is analyzed using a random walk approach in Chapter 2. The method is applied to common demand distributions such as the uniform and the normal distributions. Additionally, we adjust the method to more general cases such as non-symmetrical demands, unbalanced systems, and higher-degree chains. Chapter 3 will investigate the impact of the response dimension on the performance of flexible systems. For the symmetric system, we characterize the chaining efficiency for all relevant response levels and show that in the worst case, chaining still provides non-negligible benefits. The trade-off between system range and system response, and the complementary nature of upgrading system response and reducing demand variability are also discussed. In Chapter 4, we make a case for the often ignored third chain in terms of partial production postponement and supply disruptions. Section 4.1 shows that the 2-chain may not be enough under partial postponement, but the 3-chain can make up for most of the postpone-

ment loss. The right mix of flexibility and postponement is also explored. Section 4.2 then looks at systems subject to supply disruptions in terms of fragility, flexibility and capacity. The idea that the third chain can be a positive element under supply disruptions is also proposed and illustrated. Finally, Chapter 5 concludes with a summary of results and plans for future research.

# 2. ASYMPTOTIC CHAINING EFFICIENCY

In this chapter, we examine the effect of increasing system size on the performance of the chaining strategy vis-à-vis the full flexibility system. To this end, we consider the following simple example. Suppose each plant has a capacity of $C_j = 100$ units for each $j$, and each product consumes one unit of capacity and has an expected demand of $D_i = 100$ units for each $i$. Note that the (mean) demand and supply are balanced and identical in this case. We assume further that the demand is normally distributed with a standard deviation of 33 units (so that the probability of negative demand is negligible).

We then simulate the expected **system sales** for the dedicated system, the chaining system, and the full flexibility system. We first observe the demand realizations and then we determine how much of each product is produced by each plant in order to maximize total sales. This boils down to solving the Maximum Flow Model introduced in Section 1.3.1.

$$
\begin{aligned}
Z^*_{\mathcal{G}(n)}(\mathbf{D}) = \quad & \max \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij} \\
s.t. \quad & \sum_{j=1}^{n} x_{ij} \le D_i \ \forall i = 1, 2, \ldots n; \\
& \sum_{i=1}^{n} x_{ij} \le C_j \ \forall j = 1, 2, \ldots n; \\
& x_{ij} \ge 0 \ \forall \ i, j = 1, \ldots, n, \\
& x_{ij} = 0 \ \ \forall \ (i,j) \notin \mathcal{G}(n).
\end{aligned}
$$

We solve the above problem for each random realization of the demand vector $\mathbf{D}$. Table 2.1 shows the expected performance of the different structures over the random demand, as $n$ varies. The expected sales ratio of chaining to full flexibility and the chaining efficiency (expected benefits ratio of chaining to full flexibility) are also tabulated.

| System | Expected Sales | | | Ratios | |
|---|---|---|---|---|---|
| Size $n$ | $\mathcal{D}(n)$ | $\mathcal{C}(n)$ | $\mathcal{F}(n)$ | $SR(\mathcal{C}(n))$ | $CE(n)$ |
| 10 | 864.47 | 949.36 | 955.14 | 99.39% | 93.62% |
| 15 | 1297.51 | 1434.44 | 1447.00 | 99.13% | 91.59% |
| 20 | 1728.52 | 1915.78 | 1938.93 | 98.81% | 89.00% |
| 25 | 2179.81 | 2401.94 | 2441.73 | 98.37% | 84.81% |
| 30 | 2601.84 | 2871.06 | 2929.84 | 97.99% | 82.08% |
| 35 | 3044.48 | 3352.66 | 3430.70 | 97.73% | 79.79% |
| 40 | 3469.06 | 3807.16 | 3905.48 | 97.48% | 77.47% |

*Tab. 2.1:* Expected Sales Ratio and Chaining Efficiency as System Size Increases

For small $n$ (say $n = 10$), our simulation shows that the expected sales, essentially the maximum flow, in the three systems are 864.47, 949.36, and 955.14, respectively. This demonstrates that chaining already achieves most

(99.39%) of the expected sales of full flexibility as well as most (93.62%) of the benefits of full flexibility. As the system expands, the performance of chaining deteriorates slightly, but still at an impressive sales ratio of 97.48% and chaining efficiency of 77.47% for $n = 40$. A natural question is how well the chaining structure performs as $n$ increases to infinity. Such asymptotic analysis is important given today's growing manufacturing and service networks. This study also complements existing literature which is largely simulation-based and thus confined only to small or moderate size systems, by supplying a lower bound on the actual performance of large finite systems. The rest of this chapter is devoted to developing an analytical method that captures the asymptotic chaining performance for general demand distributions.

## 2.1 The Basic Model

Consider the case where there is an equal number of plants and products, with a (fixed) supply and (mean) demand of $\mu$ for each one; that is, the identical and balanced case. We further assume that all products have an independent and identically distributed demand $D_i$ which follows a symmetrical distribution around its mean $\mathrm{E}[D_i] = \mu$.[1] Since demand cannot be negative, we assume that $D_i \in [0, 2\mu]$ for all demand realizations. Let

[1] Our technique can be modified to handle cases when mean demand does not equal plant capacity and when the demand is not symmetrical about the mean. For the moment, we focus on the identical and balanced case merely for ease of exposition. The generalized approach is presented in Section 2.4

$\mathbf{D} = (D_1, \ldots, D_n)$ denote the demand of the $n$ products. Let

$$MF(\mathcal{G}(n), \mathbf{D})$$

denote the maximum amount of production supported by the structure $\mathcal{G}(n)$ in the system (obtained by solving the Maximum Flow Model $Z^*_{\mathcal{G}(n)}(\mathbf{D})$ in Section 1.3.1). For the dedicated and the full flexibility systems, it is easy to see that

$$MF(\mathcal{D}(n), \mathbf{D}) = \sum_{i=1}^{n} \min(\mu, D_i) = \sum_{i=1}^{n} \Big(\mu - (\mu - D_i)^+\Big),$$

and

$$MF(\mathcal{F}(n), \mathbf{D}) = \min\Big(n\mu, \sum_{i=1}^{n} D_i\Big) = n\mu + \min\Big(0, \sum_{i=1}^{n} D_i - n\mu\Big).$$

As demands are independent and bounded, by the Central Limit Theorem,

$$E\left[\min\Big(0, \sum_{i=1}^{n} D_i - n\mu\Big)\right] = \sqrt{n} E\left[\min\Big(0, \frac{\sum_{i=1}^{n}(D_i - \mu)}{\sqrt{n}}\Big)\right] \sim O(\sqrt{n}).$$

We are interested in comparing the performance of the long chain, vis-à-vis the full flexibility system. In particular, we want to evaluate the asymptotic sales ratio of chaining to full flexibility introduced in Section 1.3.2.

$$ASR(\mathcal{C}(\infty)) = \lim_{n\to\infty} \frac{\mathrm{E}[MF(\mathcal{C}(n), \mathbf{D})]}{\mathrm{E}[MF(\mathcal{F}(n), \mathbf{D})]},$$

As well as tracking the above ratio, we would also like to track the asymptotic chaining efficiency which measures the improvement of the chaining structure over the dedicated system. This refinement is useful, as it rules out those cases where the dedicated system is already as good as the full flexibility system. In fact, for the dedicated system, it is easy to show that

$$ASR(\mathcal{D}(\infty)) = \frac{\mu - E[(\mu - D_i)^+]}{\mu}.$$

By our assumption, $E[(\mu - D_i)^+] \leq \mu/2$. Hence, we already have

$$ASR(\mathcal{D}(\infty)) \geq 1/2.$$

We recall, from Section 1.3.2, the definition of chaining efficiency (or flexibility efficiency of chaining), which is just the expected benefits ratio of chaining to full flexibility.

$$\begin{aligned} CE(n) &\triangleq \frac{\mathrm{E}[MF(\mathcal{C}(n), \mathbf{D})] - \mathrm{E}[MF(\mathcal{D}(n), \mathbf{D})]}{\mathrm{E}[MF(\mathcal{F}(n), \mathbf{D})] - \mathrm{E}[MF(\mathcal{D}(n), \mathbf{D})]} \\ &= \frac{\mathrm{E}[MF(\mathcal{C}(n), \mathbf{D})] - n\mu + n\mathrm{E}[(\mu - D_i)^+]}{n\mathrm{E}[(\mu - D_i)^+] - O(\sqrt{n})}. \end{aligned}$$

Our interest is to characterize the ACE or asymptotic value of the chaining efficiency, where

$$ACE \triangleq \lim_{n \to \infty} CE(n).$$

It follows that

$$ACE = \frac{\mathrm{E}[(D_i - \mu)^+] + \lim_{n\to\infty} \frac{1}{n}\mathrm{E}[MF(\mathcal{C}(n), \mathbf{D})] - \mu}{\mathrm{E}[(D_i - \mu)^+]} \tag{2.1}$$

Hence, our focus from here onward is to find $\frac{1}{n}\mathrm{E}[MF(\mathcal{C}(n), \mathbf{D})]$. Once the asymptotic chaining efficiency is obtained, the asymptotic sales ratio of chaining to full flexibility can be easily computed as follows.

$$\begin{aligned} ASR(\mathcal{C}(\infty)) &= \lim_{n\to\infty} \frac{\mathrm{E}[MF(\mathcal{C}(n), \mathbf{D})]}{\mathrm{E}[MF(\mathcal{F}(n), \mathbf{D})]} \\ &= \mathrm{ACE} + (1 - \mathrm{ACE})\left(\lim_{n\to\infty} \frac{\mathrm{E}[MF(\mathcal{D}(n), \mathbf{D})]}{\mathrm{E}[MF(\mathcal{F}(n), \mathbf{D})]}\right) \\ &= \mathrm{ACE} + (1 - \mathrm{ACE})\left(\frac{\mu - E[(\mu - D_i)^+]}{\mu}\right) \end{aligned}$$

As the flow on each arc is bounded by $\mu$, we can delete a link from the chain $\mathcal{C}(n)$, to obtain $\mathcal{P}(n)$, without affecting the asymptotic performances of the two structures. In fact, we have the following lemma which states that the long path is asymptotically equivalent to the long chain.

**Lemma 1.**

$$\lim_{n\to\infty} \frac{E[MF(\mathcal{P}(n), \mathbf{D})]}{n} = \lim_{n\to\infty} \frac{E[MF(\mathcal{C}(n), \mathbf{D})]}{n}$$

We thus focus on finding the maximum flow on the path structure $\mathcal{P}(n)$, rather than the chain $\mathcal{C}(n)$.

## 2.2 The Random Walk Approach

For ease of exposition, we let the arc linking demand node $i$ to supply node $i$ denote the "primary" arc, and the arc linking demand node $i$ to supply node $i+1$ the "secondary" arc. We delete the arc from demand node $n$ to supply node 1 to obtain the path $\mathcal{P}(n)$. The maximum flow on the long path $\mathcal{P}(n)$ can be determined in a greedy fashion: first, satisfy the demand $D_1$ using whatever primary capacity, which is provided by the primary arc, that is available (i.e., $\mu$ units), then using as much secondary capacity, which is provided by the secondary arc, as needed (i.e., another $\mu$ units). Next, based on the level of capacity remaining, satisfy the demand $D_i$ using the primary and secondary capacities, with $i$ ranging from 2 to $n$, in that order. Note that this may (or may not) reduce the primary capacity of the next product demand, which may then need to rely more on secondary capacity. The amount of max flow obtained in this greedy fashion is a random variable, depending on the values of $D_i$.

To present this greedy approach formally and to facilitate our analysis, we let $T_i$ denote the amount of primary capacity left for product $i$ and let $S_i$ denote the amount of secondary capacity consumed by product $i$, after demands for products 1 to $i-1$ have been satisfied using the greedy method. Therefore, $T_i = \mu - S_{i-1}$ and we set $S_0 = 0$. Let TF denote the total maximum flow. Similarly, let $\text{TE} = \sum_{i=1}^{n} D_i - \text{TF}$ denote the difference between the total demand and the total flow; that is, the total unmet demand. This

implies

$$\frac{1}{n}E[\mathrm{TE}] = \mu - \frac{1}{n}E[\mathrm{TF}]. \tag{2.2}$$

Consider step $i$ of the greedy approach, wherein $T_i$ is known before $D_i$ is observed. The greedy allocation implies

$$S_i = \min[(D_i - T_i)^+, \mu], \quad T_{i+1} = \mu - S_i, \quad \mathrm{TE} = \mathrm{TE} + [(D_i - T_i)^+ - \mu]^+$$

Taking the cases when demand is above mean and below mean, we summarize the greedy approach as follows.

**Algorithm 1.** *(Greedy Approach)*

1. *Set* $i := 1, S_0 := 0, T_1 := \mu$, *and* $TE := 0$.

2. *Observe* $D_i$.

   *If* $D_i > \mu$, *then* $S_i := \min[S_{i-1} + D_i - \mu, \mu]$, $T_{i+1} = \mu - S_i$, *and* $TE := TE + \max[D_i - T_i - \mu, 0]$

   *If* $D_i < \mu$, *then* $S_i := \max[S_{i-1} + D_i - \mu, 0]$, $T_{i+1} = \mu - S_i$, *and* $TE := TE$.

3. *If* $i = n - 1$, *then STOP.* $TE := TE + \max(D_n - T_n, 0)$. *Return TE as the minimum excess.*

   *Otherwise,* $i := i + 1$ *and go to Step 2.*

At this point, note that $\{S_i : i = 0, 1, 2, \ldots\}$ behaves much like a generalized random walk, with random step size $X_i \stackrel{\Delta}{=} D_i - \mu$ and absorbing boundaries 0 and $\mu$. The value $TE$ grows in Step 2 only when $D_i - T_i > \mu$,; that is, when $S_i = \min(D_i + S_{i-1} - \mu, \mu) = \mu$. We call this quantity $(X_i - T_i)$ the level of **overshoot at the upper boundary**. Note that $(X_i - T_i) = D_i - T_i - \mu$.

In Step 2 of the greedy algorithm, when $D_i < \mu$, it is possible that $S_{i-1} + D_i - \mu < 0$. We call this amount $(-S_{i-1} - X_i)$ the level of **overshoot at the lower boundary**. Note that we do not account for overshoot at the lower boundary while keeping track of $TE$ in the greedy algorithm.

The random walk starts initially at $S_0 = 0$, the lower boundary. It gets trapped at the lower boundary whenever $X_i < 0$, and escapes only when $X_i > 0$. An interesting phenomenon happens when the random walk hits the upper boundary - the walk gets trapped at the upper boundary whenever $X_i > 0$, and it escapes from the upper boundary only when $X_i < 0$.

Let

$$\tau \stackrel{\Delta}{=} \inf \{n : S_n = \mu,\ n \geq 1\}$$

denote the stopping time when the walk first hits the upper boundary. We can re-start the random walk from the lower boundary at time $\tau$: interchange the roles of the upper and lower boundaries, and let

$$X'_i \leftarrow -X_i = \mu - D_i \ \ \forall\ i > \tau,$$

$$S'_\tau \leftarrow \mu - S_\tau = 0,$$

$$S'_i = \begin{cases} \min[S'_{i-1} + X'_i, \mu] & \text{if } X'_i > 0 \\ \max[S'_{i-1} + X'_i, 0] & \text{if } X'_i < 0 \end{cases} \quad \forall\ i > \tau.$$

Since $X_i'$ is distributed in an identical fashion to $X_i$ by symmetry of demand distribution, the random walk $S_i'$ from $S_\tau' = 0$ onwards, under the above change of co-ordinate, is identical in distribution to the earlier random walk $S_i$ starting at $S_0 = 0$.

Note that the way we account for TE changes under this new model. In the earlier walk, TE changes value only at the upper boundary, whereas in the new random walk, TE changes only when there is overshoot at the lower boundary. We repeat this process whenever the new random walk hits the upper boundary, switching back to the original random walk model. Let $\hat{S}_i$ denote the stochastic process obtained by toggling between $S_i$ and $S_i'$ in the above manner.

**Example 1.** *Figure 2.1 shows an example of a path that the random walk* $\{S_i, i = 0, 1, 2, \ldots\}$ *may traverse. Here, products 1, 3, 4, 9, and 10 have demands lower than* $\mu$*, while the rest have demands higher than* $\mu$*. We also see the walk get absorbed in the lower boundary three times and in the upper boundary once. When the walk was absorbed in the upper boundary, some unmet demands for products 6, 7, and 8 were lost. We are interested in the expected amount of such excess quantities.*

*Suppose we consider another generalized random walk* $\{\hat{S}_i, i = 0, 1, 2, \ldots\}$ *such that* $\hat{S}_0 = S_0$*, but* $\hat{S}_i$ *toggles between* $S_i' = \mu - S_i$ *and* $S_i$ *each time* $\hat{S}_i$ *hits the upper boundary. That is, the first time* $\hat{S}_i$ *hits the upper boundary, change to* $\hat{S}_i = S_i'$*; the next time, switch back to* $\hat{S}_i = S_i$*, and so on. Figure 2.2 shows*

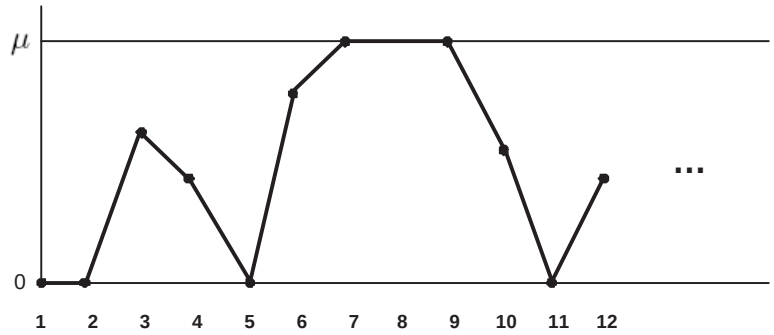


*Fig. 2.1:* Sample Path for Original Random Walk

*the equivalent sample path for the new random walk that corresponds to the sample path for the old random walk in Figure 2.1.*

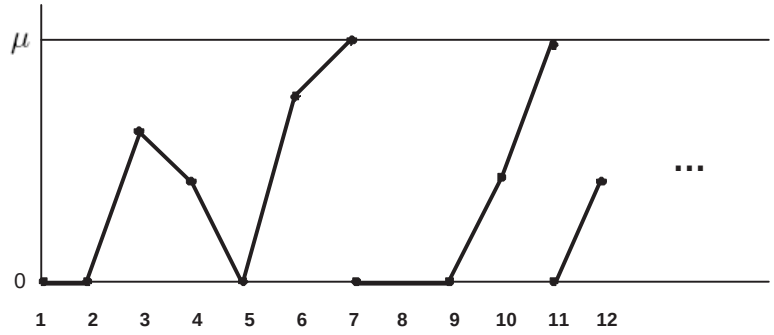


*Fig. 2.2:* Sample Path for Toggling Random Walk

*Note that unmet demand is incurred at the upper boundary when $\hat{S}_i = S_i$, but at the lower boundary when $\hat{S}_i = S'_i$. For example, in Figure 2.2, we easily verify that indeed, unmet demands are incurred for products 6, 7, and 8.*

Although it is possible to work on $\{S_i, i = 0, 1, 2, \ldots\}$, the transformation to $\{\hat{S}_i, i = 0, 1, 2, \ldots\}$ provides a more convenient formulation. In particular, $\{\hat{S}_i, i = 0, 1, 2, \ldots\}$ turns out to be a regenerative process when-

ever the random walk hits the upper boundary - the process regenerates and its continuation is a probabilistic replica of the original process starting at step 1 again.

Because all regenerating cycles are probabilistically identical, it suffices to examine the characteristics of one cycle for the purpose of asymptotic analysis. Some of these relevant characteristics are

- Cycle Duration $\tau$: the length of each regenerative cycle. Recall that

$$\tau \stackrel{\Delta}{=} \inf\{n : S_n = \mu,\ n \geq 1, S_0 = 0\}.$$

- Cycle Overshoot $\psi$: the amount of overshoots at both the lower **and** upper boundaries in each cycle.

$$\psi \stackrel{\Delta}{=} \sum_{i=1}^{\tau} \Big( (S_i - S_{i-1} - X_i)\chi(X_i < 0) + (S_{i-1} + X_i - S_i)\chi(X_i > 0) \Big),$$

  where $\chi(\cdot)$ denote the indicator function.

Note that $\psi$ can be decomposed into two components, with $\psi = \psi_L + \psi_U$, where

$$\psi_L \stackrel{\Delta}{=} \sum_{i=1}^{\tau} \Big( (S_i - S_{i-1} - X_i)\chi(X_i < 0) \Big),$$

and

$$\psi_U \stackrel{\Delta}{=} \sum_{i=1}^{\tau} \Big( (S_{i-1} + X_i - S_i)\chi(X_i > 0) \Big).$$

Consider a renewal process $\{N(t) : t \geq 0\}$, having i.i.d. inter-arrival time $Y_i$ with $Y_i \sim \tau$ for all $i$. The reward $R_i$ obtained at the $i$th renewal is

$\psi_L$ if $i$ is even, and is $\psi_U$ if $i$ is odd. Note that from (2.2),

$$\sum_{i=1}^{n} D_i - \sum_{i=1}^{N(n)+1} R_i \leq MF(\mathcal{P}_n, \mathbf{D}) \leq \sum_{i=1}^{n} D_i - \sum_{i=1}^{N(n)} R_i. \tag{2.3}$$

Because $\hat{S}_i$ toggles alternately between $S_i$ and $S'_i$ and by the renewal reward theorem,

$$\lim_{n\to\infty} \frac{\mathrm{E}[\sum_{i=1}^{N(n)} R_i]}{n} = \frac{E[\psi_L] + E[\psi_U]}{2} \frac{1}{E[\tau]}.$$

Hence, taking the limit in (2.3) obtains

**Theorem 1.**

$$\lim_{n\leftarrow\infty} \frac{\mathrm{E}[MF(\mathcal{P}(n), \mathbf{D})]}{n} = \mu - \frac{\mathrm{E}[\psi]/2}{\mathrm{E}[\tau]}.$$

For any discrete demand distribution symmetrical around the mean $\Delta$, the parameters $E[\psi]$ and $E[\tau]$ can be obtained by solving a system of linear equations. We represent the distribution as follows.

$$\text{support}\{D_i\} = \{0, 1, \ldots, \Delta, \ldots, 2\Delta - 1, 2\Delta\}$$

Let

$$P_x = \text{Prob}(D_i = \Delta + x), \quad \forall x = -\Delta, -\Delta + 1, \ldots, \Delta - 1, \Delta$$

and WLOG[2],

$$P_x = P_{-x} > 0, \quad P_0 = 0$$

[2] Suppose $P_0 > 0$. Let $P'_0 = 0, P'_x = \frac{P_x}{1-P_0}, \forall x \neq 0$. It follows that $\mathrm{E}[\tau'] = (1 - P_0)\mathrm{E}[\tau]$ and $\mathrm{E}[\psi'] = (1 - P_0)\mathrm{E}[\psi]$

Define the stopping time if the random walk started at $x$.

$$\tau_x \stackrel{\Delta}{=} \inf\{n : S_n = \mu,\ n \geq 1, S_0 = x\}$$

Clearly, $\tau = \tau_0$, and $\tau_\Delta = 0$. Conditioning on the next move,

$$\mathrm{E}[\tau_x] = 1 + \sum_{j=1}^{\Delta-1} \mathrm{E}[\tau_j] P_{j-x} + \mathrm{E}[\tau_0] \sum_{j=x}^{\Delta} P_j, \quad \forall x = 0, 1, \ldots, \Delta - 1 \qquad (2.4)$$

We can obtain $E[\tau] = E[\tau_0]$ by solving the system of equation (2.4).

Similarly, given $S_0 = x$, we define the overshoot as

$$\psi_x \stackrel{\Delta}{=} \sum_{i=1}^{\tau_x} \Big( (S_i - S_{i-1} - X_i)\chi(X_i < 0) + (S_{i-1} + X_i - S_i)\chi(X_i > 0) \Big)$$

Obviously, $\psi = \psi_0$ and $\psi_\Delta = 0$. Conditioning on the next move,

$$\mathrm{E}[\psi_x] = r_x + \sum_{j=1}^{\Delta-1} \mathrm{E}[\psi_j] P_{j-x} + \mathrm{E}[\psi_0] \sum_{j=x}^{\Delta} P_j, \quad \forall x = 0, 1, \ldots, \Delta - 1 \qquad (2.5)$$

where

$$r_x \ = \ \sum_{j=\Delta}^{\Delta+x} (j - \Delta) P_{j-x} + \sum_{j=x}^{\Delta} (j - x) P_j, \quad \forall x = 0, 1, \ldots, \Delta - 1$$

We can obtain $E[\psi] = E[\psi_0]$ by solving the system of equation (2.5).

By Theorem 1 and the definition of $ACE$, we have

**Theorem 2.** *The asymptotic chaining efficiency can be uniquely obtained as follows.*

$$ACE = 1 - \frac{\mathrm{E}[\psi_0]}{2\mathrm{E}[\tau_0]\mathrm{E}[(D_i - \Delta)^+]}$$

*where* $\mathrm{E}[\psi_0]$ *and* $\mathrm{E}[\tau_0]$ *come from the solutions to linear systems (2.4) and (2.5), respectively.*

**Proof.** We show first the uniqueness of the solutions to (2.4) and (2.5). Observe that (2.4) and (2.5) have the same homogeneous system. Since $\sum_{j=1}^{\Delta-1} P_{j-x} + \sum_{j=x}^{\Delta} P_j < 1$, the associated matrix is strictly diagonally dominated, hence nonsingular.

Now, from (2.1), Lemma 1 and Theorem 1,

$$\begin{aligned} ACE &= \frac{\mathrm{E}[(D_i - \Delta)^+] + \lim_{n\to\infty} \frac{1}{n}\mathrm{E}[MF(\mathcal{P}(n), \mathbf{D})] - \Delta}{\mathrm{E}[(D_i - \Delta)^+]} \\ &= 1 - \frac{\mathrm{E}[\psi_0]}{2\mathrm{E}[\tau_0]\mathrm{E}[(D_i - \Delta)^+]} \end{aligned}$$

■

Furthermore, ACE is invariant over the scale of the demand.

**Corollary 1.** *Suppose* $D_i' \sim cD_i, c > 0$. *Then,* $ACE' = ACE$.

**Proof.** It is easy to see that $\mathrm{E}[\tau_0'] = \mathrm{E}[\tau_0]$, $\mathrm{E}[\psi_0'] = c\mathrm{E}[\psi_0]$, and $\mathrm{E}[(D_i' - c\Delta)^+] = c\mathrm{E}[(D_i - \Delta)^+]$, from which the result follows. ■

This gives rise to an efficient method to determine the asymptotic efficiency of the 2-chain. When demand follows a discrete distribution, Theorem 2 and Corollary 1 can be directly applied. On the other hand, when demand follows a continuous distribution from 0 to $2\mu$, the above results can still be used to approximate the asymptotic chaining performance. This is done by discretizing the distribution into $2\Delta + 1$ equally spaced demand points from 0 to $2\mu$. Obviously, the more discrete points used, the better the approximation.

## 2.3 Applications

### 2.3.1 Two-Point Distribution

When demand $D_i = 0$ or $2\mu$ with equal probability, then it is easy to see that

$$E[\psi] = \mu, \;\; E[\tau] = 2 \quad \Rightarrow ACE = 0.5$$

Furthermore, since $E[(\mu - D_i)^+] = \mu/2$,

$$\begin{aligned} ASR(\mathcal{C}(\infty)) &= ACE + (1 - ACE)\left(\frac{\mu - \mathrm{E}[(\mu - D_i)^+]}{\mu}\right) \\ &= 0.5 + (1 - 0.5)(1 - 0.5) = 0.75 \end{aligned}$$

Thus, the chaining strategy achieves only 75% of the efficiency of the full flexibility system. This poor performance stems in part from the large variability in the demand distribution.

### 2.3.2 Uniform Distribution

Suppose demand $D_i = 0, 1, \ldots, \Delta - 1, \Delta + 1, \ldots, 2\Delta - 1, 2\Delta$ with equal probability; that is,

$$P_x = \frac{1}{2\Delta}, \quad \forall x = 1, \ldots, \Delta - 1, \Delta$$

It can be shown that

$$\mathrm{E}[\tau_0] = \frac{4\Delta(2\Delta + 1)}{(\Delta + 2)(\Delta + 1)}, \quad \mathrm{E}[\psi_0] = \frac{\Delta(5\Delta + 4)}{3(\Delta + 2)}, \quad \mathrm{E}[D_i - \Delta]^+ = \frac{\Delta + 1}{4}$$

Hence,

$$ACE = \frac{7\Delta + 2}{12\Delta + 6}.$$

Furthermore, since $E[(\Delta - D_i)^+] = (\Delta + 1)/4$,

$$\begin{aligned} ASR(\mathcal{C}(\infty)) &= \frac{7\Delta + 2}{12\Delta + 6} + \frac{5\Delta + 4}{12\Delta + 6} \times \left(0.75 - \frac{1}{4\Delta}\right) \\ &= \frac{43\Delta^2 + 15\Delta - 4}{48\Delta^2 + 24\Delta} \end{aligned}$$

When demand $D_i$ is uniformly distributed over $[0, 2\mu]$, we can obtain the ACE by first discretizing the interval into $2\Delta + 1$ demand points, then taking the limit as $\Delta \to \infty$. Hence

$$ACE = \lim_{\Delta \to \infty} \frac{7\Delta + 2}{12\Delta + 6} = \frac{7}{12} \approx 58.33\%$$

and

$$ASR(\mathcal{C}(\infty)) = \lim_{\Delta\to\infty} \frac{43\Delta^2 + 15\Delta - 4}{48\Delta^2 + 24\Delta} = \frac{43}{48} \approx 89.58\%$$

Note that in this case, the value of the expected max flow in the 2-chain is already around 89.6% of the expected max flow in the full flexibility system!

### 2.3.3 Normal Distribution

Suppose demand $D_i \sim N(\mu, \sigma)$. Then, we can likewise approximate the value of $ACE$ by discretization. Moreover, Corollary 1 implies that for a fixed coefficient of variation, the $ACE$ is independent of the actual magnitudes of $\mu$ and $\sigma$. Table 2.2 summarizes how the $ACE$ values change with respect to the discretization level $\Delta$ and the coefficient of variation $CV$.

| | Coefficient of Variation (CV) | | | | | | |
|---|---|---|---|---|---|---|---|
| $\Delta$ | 0.33 | 0.31 | 0.29 | 0.27 | 0.25 | 0.23 | 0.21 |
| 2 | 0.6452 | 0.6509 | 0.6559 | 0.6599 | 0.6629 | 0.6649 | 0.6660 |
| 4 | 0.6895 | 0.7007 | 0.7124 | 0.7244 | 0.7365 | 0.7486 | 0.7604 |
| 6 | 0.6970 | 0.7090 | 0.7216 | 0.7348 | 0.7484 | 0.7623 | 0.7765 |
| 8 | 0.6997 | 0.7119 | 0.7248 | 0.7383 | 0.7524 | 0.7669 | 0.7819 |
| 10 | 0.7010 | 0.7133 | 0.7263 | 0.7399 | 0.7542 | 0.7690 | 0.7843 |
| 12 | 0.7017 | 0.7140 | 0.7271 | 0.7408 | 0.7552 | 0.7701 | 0.7856 |
| 14 | 0.7022 | 0.7145 | 0.7275 | 0.7413 | 0.7558 | 0.7708 | 0.7864 |

*Tab. 2.2:* Asymptotic Chaining Efficiency for Various Levels of Discretization and Demand Uncertainty

To handle negative demand, we truncated the distribution to have finite support $[0, 2\mu]$. Because the resulting distribution remains symmetric around the mean, our random walk approach works. For the results reported in Table 2.2, we considered $CV \leq 0.33$, which implies negligible probability of negative demand. Therefore, the accuracy loss is also negligible. For higher

values of $CV$, it is still reasonable in many cases to truncate the distribution in the same manner and the approach still works. Nevertheless, if a more realistic truncation results in a non-symmetrical demand distribution, then we have to use an extended version of our random walk approach. We defer this discussion to Section 2.4.

The table also shows that as we increase the number of demand points, the approximation becomes finer. More importantly, the value of ACE decreases in the coefficient of variation. This is because as relative uncertainty decreases, the need for any form of flexibility is reduced, thus improving the value of the 2-chain relative to full flexibility.

We tabulate next the ratio of the expected sales from the chaining structure and the full flexibility system in Table 2.3. Interestingly, even with a CV of 0.33, the expected sales under the chaining structure are already close to 96% of the full flexibility system.

| | Coefficient of Variation (CV) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0.33 | 0.31 | 0.29 | 0.27 | 0.25 | 0.23 | 0.21 |
| $ASR(\mathcal{C}(\infty))$ | 0.9614 | 0.9650 | 0.9687 | 0.9723 | 0.9758 | 0.9791 | 0.9823 |

*Tab. 2.3:* Asymptotic Sales Ratio for Various Levels of Demand Uncertainty

## 2.4 Extensions

The proposed method works so long as all products have the same demand distribution and all plants have the same capacity, even if the system is unbalanced (i.e., capacity not equal to mean demand) and the demand distribution is not symmetrical around the mean.

### 2.4.1 New Random Walk: Alternating Renewal Process

Consider the case when demand is not symmetrical around its mean and expected demand is not equal to fixed capacity. We assume that all products have an independent and identically distributed demand $D_i$ which follows a general distribution with mean $\mathrm{E}[D_i] = \mu$. Each plant has a capacity of $C$ units. Let $\mathbf{D} = (D_1, \ldots, D_n)$ denote the demand of the $n$ products. Let

$$MF(\mathcal{G}(n), \mathbf{D})$$

denote the maximum amount of production supported by the structure $\mathcal{G}(n)$ in the system. For the dedicated and full flexibility system, it is easy to see that

$$MF(\mathcal{D}(n), \mathbf{D}) = \sum_{i=1}^{n} \min(C, D_i) = \sum_{i=1}^{n} \Big( C - (C - D_i)^+ \Big),$$

and

$$\begin{aligned} MF(\mathcal{F}(n), \mathbf{D}) &= \min\Big( nC, \sum_{i=1}^{n} D_i \Big) = nC + \min\Big( 0, \sum_{i=1}^{n} D_i - nC \Big) \\ &= \sum_{i=1}^{n} D_i - \Big( \sum_{i=1}^{n} D_i - nC \Big)^+ \end{aligned}$$

As demand is independent and bounded, by the Central Limit Theorem,

$$E\left[ \left( \sum_{i=1}^{n} D_i - nC \right)^+ \right] = \sqrt{n} E\left[ \left( \frac{\sum_{i=1}^{n} D_i - n\mu}{\sqrt{n}} + \frac{n(\mu - C)}{\sqrt{n}} \right)^+ \right] \sim O(\sqrt{n}).$$

As before, we are interested in the asymptotic sales ratio of chaining to

full flexibility.

$$ASR(\mathcal{C}(\infty)) = \lim_{n\to\infty} \frac{\mathrm{E}[MF(\mathcal{C}(n), \mathbf{D})]}{\mathrm{E}[MF(\mathcal{F}(n), \mathbf{D})]},$$

We would also like to track both the chaining efficiency and the asymptotic chaining efficiency as previously defined.

$$\begin{aligned} CE(n) &\triangleq \frac{\mathrm{E}[MF(\mathcal{C}(n), \mathbf{D})] - \mathrm{E}[MF(\mathcal{D}(n), \mathbf{D})]}{\mathrm{E}[MF(\mathcal{F}(n), \mathbf{D})] - \mathrm{E}[MF(\mathcal{D}(n), \mathbf{D})]} \\ &= \frac{\mathrm{E}[MF(\mathcal{C}(n), \mathbf{D})] - nC + n\mathrm{E}[(C - D_i)^+]}{n\mu - O(\sqrt{n}) - nC + n\mathrm{E}[(C - D_i)^+]} \\ &= \frac{\mathrm{E}[MF(\mathcal{C}(n), \mathbf{D})] + n\mathrm{E}[(D_i - C)^+] - n\mu}{n\mathrm{E}[(D_i - C)^+] - O(\sqrt{n})}. \end{aligned}$$

and

$$ACE \triangleq \lim_{n\to\infty} CE(n).$$

It follows that

$$ACE = \frac{\lim_{n\to\infty} \frac{1}{n}\mathrm{E}[MF(\mathcal{C}(n), \mathbf{D})] + \mathrm{E}[(D_i - C)^+] - \mu}{\mathrm{E}[(D_i - C)^+]}$$

Using an alternating renewal process, we extend the random walk method developed for the symmetric case as follows.

$$\lim_{n\to\infty} \frac{\mathrm{E}[MF(\mathcal{C}(n), \mathbf{D})]}{n} = \mu - \frac{\mathrm{E}[\psi_0] + \mathrm{E}[\hat{\psi}_0]}{\mathrm{E}[\tau_0] + \mathrm{E}[\hat{\tau}_0]}$$

where $\tau_i$ and $\hat{\tau}_i$ are the stopping times for the odd and even cycles, respectively, while $\psi_i$ and $\hat{\psi}_i$ denote the respective overshoots. Hence,

$$ACE = 1 - \frac{\mathrm{E}[\psi_0] + \mathrm{E}[\hat{\psi}_0]}{\mathrm{E}[\tau_0] + \mathrm{E}[\hat{\tau}_0]} \cdot \frac{1}{\mathrm{E}[D_i - C]^+}$$

Moreover,

$$ASR(\mathcal{C}(\infty)) = \frac{\mu - \frac{\mathrm{E}[\psi_0]+\mathrm{E}[\hat{\psi}_0]}{\mathrm{E}[\tau_0]+\mathrm{E}[\hat{\tau}_0]}}{\mu} = 1 - \frac{\mathrm{E}[\psi_0] + \mathrm{E}[\hat{\psi}_0]}{\mathrm{E}[\tau_0] + \mathrm{E}[\hat{\tau}_0]} \cdot \frac{1}{\mu}$$

### 2.4.2 Example: Non-symmetrical Demand

In the case of non-symmetrical demand, the odd and the even cycles will have different stopping times and overshoots. To demonstrate, we consider the following example.

$$D_i = \begin{cases} 0, & \text{w.p. } 0.3 \\ 2, & \text{w.p. } 0.1 \\ 4, & \text{w.p. } 0.4 \\ 8, & \text{w.p. } 0.2 \end{cases} \quad \text{and} \quad C_i = 5$$

Note that $D_i$ is not symmetrical about the mean. To obtain the asymptotic chaining efficiency and the asymptotic sales ratio for this scenario, we solve

the following systems of linear equations.

$$\begin{bmatrix} 0.2 & 0 & 0 & -0.2 & 0 \\ -0.8 & 1 & 0 & 0 & -0.2 \\ -0.4 & -0.4 & 1 & 0 & 0 \\ -0.4 & 0 & -0.4 & 1 & 0 \\ -0.3 & -0.1 & 0 & -0.4 & 1 \end{bmatrix} \begin{bmatrix} \mathrm{E}[\tau_0] & \mathrm{E}[\psi_0] \\ \mathrm{E}[\tau_1] & \mathrm{E}[\psi_1] \\ \mathrm{E}[\tau_2] & \mathrm{E}[\psi_2] \\ \mathrm{E}[\tau_3] & \mathrm{E}[\psi_3] \\ \mathrm{E}[\tau_4] & \mathrm{E}[\psi_4] \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0.2 \\ 1 & 0.4 \end{bmatrix}$$

and

$$\begin{bmatrix} 0.8 & -0.4 & 0 & -0.1 & 0 \\ -0.2 & 1 & -0.4 & 0 & -0.1 \\ -0.2 & 0 & 1 & -0.4 & 0 \\ -0.2 & 0 & 0 & 1 & -0.4 \\ 0 & -0.2 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathrm{E}[\hat{\tau}_0] & \mathrm{E}[\hat{\psi}_0] \\ \mathrm{E}[\hat{\tau}_1] & \mathrm{E}[\hat{\psi}_1] \\ \mathrm{E}[\hat{\tau}_2] & \mathrm{E}[\hat{\psi}_2] \\ \mathrm{E}[\hat{\tau}_3] & \mathrm{E}[\hat{\psi}_3] \\ \mathrm{E}[\hat{\tau}_4] & \mathrm{E}[\hat{\psi}_4] \end{bmatrix} = \begin{bmatrix} 1 & 0.6 \\ 1 & 0.4 \\ 1 & 0.2 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

Hence,

$$\begin{aligned} ACE &= 1 - \frac{\mathrm{E}[\psi_0] + \mathrm{E}[\hat{\psi}_0]}{\mathrm{E}[\tau_0] + \mathrm{E}[\hat{\tau}_0]} \cdot \frac{1}{\mathrm{E}[D_i - C_i]^+} \\ &= 1 - \frac{0.7436 + 1.2377}{22.7920 + 2.8798} \cdot \frac{1}{0.6} \\ &= 0.8714 \end{aligned}$$

and

$$\begin{aligned} ASR(\mathcal{C}(\infty)) &= 1 - \frac{\mathrm{E}[\psi_0] + \mathrm{E}[\hat{\psi}_0]}{\mathrm{E}[\tau_0] + \mathrm{E}[\hat{\tau}_0]} \cdot \frac{1}{\mu} \\ &= 1 - \frac{0.7436 + 1.2377}{22.7920 + 2.8798} \cdot \frac{1}{3.4} \\ &= 0.9773 \end{aligned}$$

### 2.4.3 Example: Unbalanced System

We consider next the situation when the total supply capacity may not be the same as the total demand. Consider the case when the demands are normally distributed with mean $\mu$ and standard deviation $\sigma$ (with a CV of at most 0.33), but the capacity of each plant is $\lambda$.

When $\lambda = \mu$, we note that the absence of a safety capacity entails a fill rate of only 100(1-0.399× CV)% = 86.7% for each product, when CV=0.33. To guarantee a 97.26% fill rate for each product, a dedicated system ought to carry a safety capacity of $\sigma$ units for each product, leading to a total safety capacity of $n\sigma$ units. Since full flexibility corresponds to complete capacity pooling, we can achieve a 97.26% fill rate for the entire system with only $\sigma\sqrt{n}$ units of safety capacity. This dramatic reduction in safety stock investment performance comes about with full flexibility in the production system. We investigate the corresponding performance in the case of the chaining structure.

Tables 2.4 and 2.5 demonstrate that both $ACE$ and $ASR(\mathcal{C}(\infty))$ increase in the ratio of supply to mean demand ($\lambda/\mu$). This suggests that the balanced scenario ($\lambda = \mu$) provides a lower bound for the situation when

$\lambda > \mu$ (i.e., safety capacity scenario). Therefore, with $\lambda = \mu + \sigma/\sqrt{n}$ (i.e., $\sigma\sqrt{n}$ units of total safety capacity in the system), a chaining structure can already guarantee a fill rate of at least $97.26\% \times 96.14\% = 93.5\%$. Note that the average safety capacity per plant is decreasing in $n$. For the dedicated system to maintain this level of fill rate, the corresponding safety capacity investment is at least $0.5n\sigma$. This analysis suggests another advantage of flexibility in production planning - apart from increasing the expected sales, the flexibility strategy can also help to decrease the safety capacity investment needed to maintain a required fill-rate level. In the identical demand case, we expect that the safety capacity investment needed should decrease roughly by a factor of $O(\sqrt{n})$.

| | Coefficient of Variation (CV) | | | | | | |
|---|---|---|---|---|---|---|---|
| $\lambda/\mu$ | 0.33 | 0.31 | 0.29 | 0.27 | 0.25 | 0.23 | 0.21 |
| 0.85 | 0.2986 | 0.2810 | 0.2616 | 0.2400 | 0.2162 | 0.1904 | 0.1628 |
| 0.90 | 0.4156 | 0.4035 | 0.3892 | 0.3720 | 0.3513 | 0.3265 | 0.2972 |
| 0.95 | 0.5561 | 0.5552 | 0.5531 | 0.5492 | 0.5428 | 0.5328 | 0.5180 |
| 1.00 | 0.7037 | 0.7159 | 0.7290 | 0.7428 | 0.7574 | 0.7726 | 0.7885 |
| 1.05 | 0.8314 | 0.8510 | 0.8715 | 0.8924 | 0.9136 | 0.9345 | 0.9541 |
| 1.10 | 0.9189 | 0.9365 | 0.9529 | 0.9673 | 0.9794 | 0.9886 | 0.9947 |
| 1.15 | 0.9659 | 0.9771 | 0.9859 | 0.9923 | 0.9964 | 0.9986 | 0.9996 |

*Tab. 2.4:* Asymptotic Chaining Efficiency for Various Levels of Safety Capacity and Demand Uncertainty

### *2.4.4 Higher-degree Chains*

In the case of higher-degree chains, the extension of the random walk method is quite simple. For a $d$-chain, the corresponding $ACE_d$ can be obtained using a similar random walk with the sole exception that the upper boundary is

| | Coefficient of Variation (CV) | | | | | | |
|---|---|---|---|---|---|---|---|
| $\lambda/\mu$ | 0.33 | 0.31 | 0.29 | 0.27 | 0.25 | 0.23 | 0.21 |
| 0.85 | 0.8467 | 0.8475 | 0.8482 | 0.8489 | 0.8494 | 0.8497 | 0.8499 |
| 0.90 | 0.8912 | 0.8930 | 0.8948 | 0.8964 | 0.8978 | 0.8988 | 0.8994 |
| 0.95 | 0.9304 | 0.9335 | 0.9365 | 0.9394 | 0.9421 | 0.9444 | 0.9464 |
| 1.00 | 0.9614 | 0.9650 | 0.9687 | 0.9723 | 0.9758 | 0.9791 | 0.9823 |
| 1.05 | 0.9820 | 0.9852 | 0.9882 | 0.9909 | 0.9934 | 0.9955 | 0.9972 |
| 1.10 | 0.9930 | 0.9950 | 0.9966 | 0.9979 | 0.9988 | 0.9994 | 0.9998 |
| 1.15 | 0.9977 | 0.9986 | 0.9992 | 0.9996 | 0.9998 | 1.0000 | 1.0000 |

*Tab. 2.5:* Asymptotic Sale Ratio for Various Levels of Safety Capacity and Demand Uncertainty

stretched to $(d-1)\mu$. That is, the upper boundary is $2\mu$ for 3-chain, $3\mu$ for 4-chain, and so on.

With this new method, we can compute for $ACE_d$ and $ASR(\mathcal{C}_d(\infty))$ for different values of $d$ and for normal demand distributions with coefficient of variation (CV) values ranging from 0.21 to 0.33. The numerical results are presented in Table 2.6 and Table 2.7, respectively.

| | Coefficient of Variation (CV) | | | | | | |
|---|---|---|---|---|---|---|---|
| $d$ | 0.33 | 0.31 | 0.29 | 0.27 | 0.25 | 0.23 | 0.21 |
| 2 | 0.7046 | 0.7161 | 0.7287 | 0.7423 | 0.7567 | 0.7718 | 0.7875 |
| 3 | 0.8294 | 0.8368 | 0.8449 | 0.8536 | 0.8628 | 0.8723 | 0.8821 |
| 4 | 0.8800 | 0.8855 | 0.8914 | 0.8978 | 0.9045 | 0.9114 | 0.9184 |
| 5 | 0.9072 | 0.9116 | 0.9164 | 0.9215 | 0.9267 | 0.9321 | 0.9376 |

*Tab. 2.6:* Asymptotic Chaining Efficiency for Various Levels of Partial Flexibility and Demand Uncertainty

We observe the considerable gains in $ACE_d$ and in $ASR(\mathcal{C}_d(\infty))$ that one can obtain by upgrading from 2-chain to 3-chain. These gains are high relative to the improvements derived from 3-chain to 4-chain. Further on, the gains from 4-chain to 5-chain becomes negligible. This insight tells us that

| | Coefficient of Variation (CV) | | | | | | |
|---|---|---|---|---|---|---|---|
| $d$ | 0.33 | 0.31 | 0.29 | 0.27 | 0.25 | 0.23 | 0.21 |
| 2 | 0.9614 | 0.9650 | 0.9687 | 0.9723 | 0.9758 | 0.9791 | 0.9822 |
| 3 | 0.9777 | 0.9799 | 0.9821 | 0.9843 | 0.9863 | 0.9883 | 0.9901 |
| 4 | 0.9843 | 0.9859 | 0.9875 | 0.9890 | 0.9905 | 0.9919 | 0.9932 |
| 5 | 0.9879 | 0.9891 | 0.9904 | 0.9916 | 0.9927 | 0.9938 | 0.9948 |

*Tab. 2.7:* Asymptotic Sales Ratio for Various Levels of Partial Flexibility and Demand Uncertainty

when the system size is very large, there might be a need for more flexibility. The good news is that a third layer (or at most a fourth layer) of flexibility appears to be enough to capture most of the loss brought about by system expansion.

# 3. RANGE AND RESPONSE: DIMENSIONS OF FLEXIBILITY

The current literature on process flexibility relies on the assumption that each facility can produce any assigned product with the same efficiency. This does not take into account the possibility that facilities primarily designed to produce certain products can only serve as less efficient secondary (or back-up) production options for other products. This concern is not entirely new. In an early work, Slack [49] suggested that flexibility has two dimensions. One dimension involves the range of states a production or service system can adopt. A system is considered more flexible than another if it can take on a wider range of states, for example, make a greater variety of products. Hence, a 2-chain system is more flexible than a dedicated system. However, this property by itself does not completely describe the flexibility of a system. The ease with which the system switches from one state to another in terms of cost, time, or organizational disruption is also vital. A system that switches more quickly, smoothly, or cheaply from state to state should likewise be considered more flexible than a system that does the same at greater expense. Termed "range" and "response" in the literature [49], these two dimensions and the importance of their distinction have been acknowledged in practice

as observed in interviews with managers [49]. In fact, managers believe that differentiating between range and response has helped them articulate their flexibility needs.

To the best of our knowledge, there has been no paper in the literature that analytically examines both dimensions of process flexibility as most papers only consider range flexibility (e.g. partial flexibility versus full flexibility). One way to bring in the response dimension is to consider the setup time or the setup cost incurred when switching from producing product A to product B. Both are undesirable as setup time effectively reduces capacity whereas setup cost reduces total profits. Moreover, a fully flexible system can be expected to exhibit more production switching than a less flexible system like chaining. Modeling response this way, chaining efficiency or the performance of sparse structures can only improve as the response level deteriorates. Hence, the core models in Chapter 2 and in Chou et al. [16] are robust against such setup effects.

In this chapter, we model the response dimension in terms of production efficiencies. We distinguish between primary and secondary production, such that primary production is at least as cheap as secondary production.[1] If the cost of secondary production is high, we say that **the response is low**. Otherwise, if the cost of secondary production is low (comparable to primary production), we say that **the response is high**. In the special case when the cost of secondary production equals the cost of primary production, we

[1] Production efficiency can also be modeled in terms of production time. In that case, we can approximate increased production time by increased production cost in the sense that in order to retain the original production speed, one has to spend more on other resources like labor.

say that **the response is perfect**. To incorporate these production costs, we use an expected profit criterion for evaluating process flexibility. This criterion generalizes the expected sales criterion when system response is perfect. Once again, we analyze the **asymptotic performance** of process flexibility structures for all response levels when the system size becomes infinitely large.

## *3.1 The General Model*

We consider a firm with $n$ products and $n$ plants. Product $i$ has random demand $D_i$ whereas plant $j$ has fixed capacity $C_j$. At this stage, we do not make any assumptions on the demand distribution nor on system asymmetry. To model the response dimension, we suppose that plant $i$ is designed to produce product $i$ primarily. Only as a back-up option, it can also produce product $j \neq i$ as a secondary product but less efficiently. Each unit of product $i$ sold earns the firm $p$ dollars. Without loss of generality, we ignore the goodwill cost associated with unsatisfied demand.[2] Now if this unit of demand is produced by its primary plant $i$, the production cost is $c_p$. On the other hand, this same product produced by a secondary plant $j \neq i$ costs the firm at least as much at $c_s \geq c_p$. We call $c_p$ and $c_s$ the costs of primary and secondary production, respectively. To avoid triviality, we assume $c_s < p$. We can then use the cost parameter $c_s$ to capture the system response level as summarized in the table below.

[2] In the case where the firm incurs a goodwill cost of $g$ for every unit of unsatisfied demand, we add the goodwill cost to the unit revenue and get the imputed price $\bar{p} = p + g$. Replacing $p$ with $\bar{p}$, we use the same model and get the same analytical results.

| Response | Value of $c_s$ |
|---|---|
| Low | $c_s \geq \frac{1}{2}(p + c_p)$ |
| High | $c_p < c_s < \frac{1}{2}(p + c_p)$ |
| Perfect | $c_s = c_p$ |

*Tab. 3.1:* Summary of System Response Levels

As before, our goal is to characterize how a given flexibility configuration performs relative to full flexibility. We use the edge set $\mathcal{G}(n)$ of a bipartite graph to represent the flexibility configuration under consideration.

Let $\mathbf{D} = (D_1, \ldots, D_n)$ be the demand vector and let $x_{ij}$ be the number of units of product $i$ produced by facility $j$. Given a system response level $c_s$, a flexibility configuration $\mathcal{G}(n)$ and demand realization $\mathbf{D}$, the task boils down to solving the following Maximum Profit Model introduced in Section 1.3.1.

$$\begin{aligned} \Pi^*_{\mathcal{G}(n)}(\mathbf{D}, c_s) = \quad \max \quad & (p - c_p) \sum_{i=1}^{n} x_{ii} + (p - c_s) \sum_{i=1}^{n} \sum_{j \neq i} x_{ij} \qquad (3.1) \\ \text{s.t.} \quad & \sum_{j=1}^{n} x_{ij} \leq D_i \quad \forall\ i = 1, \ldots, n \\ & \sum_{i=1}^{n} x_{ij} \leq C_j \quad \forall\ j = 1, \ldots, n \\ & x_{ij} \geq 0 \quad \forall\ i, j = 1, \ldots, n \\ & x_{ij} = 0 \quad \forall\ (i, j) \notin \mathcal{G}(n) \end{aligned}$$

We want to study the flexibility efficiency measure introduced in Section 1.3.2, which captures the incremental benefits of any flexibility structure over the default dedicated system, relative to the best system, which is the full

flexibility system.

$$FE(\mathcal{G}(n), c_s) = \frac{\mathrm{E}[\Pi^*_{\mathcal{G}(n)}(\mathbf{D}, c_s)] - \mathrm{E}[\Pi^*_{\mathcal{D}(n)}(\mathbf{D}, c_s)]}{\mathrm{E}[\Pi^*_{\mathcal{F}(n)}(\mathbf{D}, c_s)] - \mathrm{E}[\Pi^*_{\mathcal{D}(n)}(\mathbf{D}, c_s)]} \tag{3.2}$$

It can be shown that for any flexibility structure, its flexibility efficiency may worsen as system response worsens (see Theorem 3). It follows that this profit-based flexibility efficiency is never more than the sales-based flexibility efficiency considered in Jordan and Graves [32] and in Chapter 2 of this thesis. Hence, it is a more conservative flexibility measure.

Now, observe that regardless of the value of $c_s$, it is easy to derive the optimal production allocations for both the dedicated system and the full flexibility systems. For the dedicated system, each facility can only produce its primary product. Therefore, the optimal allocation is for each facility to produce as many units of its primary product as possible. On the other hand, for the full flexibility system, any facility can produce any product. Thus, it is optimal for each facility to produce as many units of its primary product as possible, and *only* thereafter, use its extra capacity, if any, to produce the extra demand, if any, of *any* other product. We have the following optimal allocation for the dedicated system:

$$\begin{aligned} x^*_{ii} &= \min(D_i, C_i) \quad \forall\ i \\ x^*_{ij} &= 0 \quad \forall\ i, \forall\ j \neq i \end{aligned} \tag{3.3}$$

and the following optimal allocation for the fully flexible system:

$$\begin{aligned} x^*_{ii} &= \min(D_i, C_i) \quad \forall\ i \\ \textstyle\sum_i \sum_{j\neq i} x^*_{ij} &= \min[\textstyle\sum_i (D_i - C_i)^+, \sum_i (C_i - D_i)^+] \\ &= \min(\textstyle\sum_i D_i, \sum_i C_i) - \sum_i \min(D_i, C_i) \end{aligned} \tag{3.4}$$

The last equation can be interpreted intuitively as total secondary production equaling total production minus total primary production.

Observe that the arguments of $FE(\mathcal{G}(n), c_s)$ completely capture the dimensions of process flexibility, as $|\mathcal{G}|$ and $c_s$ respectively represent the range and response levels. For $\mathcal{G}_1(n) \subseteq \mathcal{G}_2(n)$, it is easy to see that $FE(\mathcal{G}_1(n), c_s) \leq FE(\mathcal{G}_2(n), c_s)$ since $\mathcal{G}_2(n)$ has a larger feasible region. This means that upgrading system range improves system performance. The same can also be said about upgrading system response as shown in the following theorem.

**Theorem 3.** *For a fixed flexibility structure $\mathcal{G}(n)$, such that $\mathcal{D}(n) \subseteq \mathcal{G}(n) \subseteq \mathcal{F}(n)$, its flexibility efficiency $FE(\mathcal{G}(n), c_s)$ is non-increasing in $c_s$ over the interval $[c_p, p)$.*

*Proof.* Consider $c_s > c'_s$. For a fixed structure $\mathcal{G}(n)$ and a demand realization $\mathbf{D}$, we let $X_P = \sum_{i=1}^n x^*_{ii}$ and $X_S = \sum_{i=1}^n \sum_{j\neq i} x^*_{ij}$ be the optimal primary and secondary production, respectively, when secondary production cost is $c_s$. Similarly, $X'_P$ and $X'_S$ are the optimal primary and secondary production when secondary production cost is $c'_s$. From model (3.1), equations (3.3) and

(3.4), and the definition of flexibility efficiency (3.2), we obtain the following:

$$FE(\mathcal{G}(n), c_s) = \frac{\mathrm{E}[(p-c_s)X_S - (p-c_p)(\sum_i \min(D_i, C_i) - X_P)]}{\mathrm{E}[(p-c_s)(\min(\sum_i D_i, \sum_i C_i) - \sum_i \min(D_i, C_i))]}$$

Hence,

$$\begin{aligned} FE(\mathcal{G}(n), c_s) &= \frac{\mathrm{E}[X_S - \frac{p-c_p}{p-c_s}(\sum_i \min(D_i, C_i) - X_P)]}{\mathrm{E}[((\min(\sum_i D_i, \sum_i C_i) - \sum_i \min(D_i, C_i))]} \\ &\leq \frac{\mathrm{E}[X_S - \frac{p-c_p}{p-c'_s}(\sum_i \min(D_i, C_i) - X_P)]}{\mathrm{E}[((\min(\sum_i D_i, \sum_i C_i) - \sum_i \min(D_i, C_i))]} \\ &\leq \frac{\mathrm{E}[(p-c'_s)X'_S - (p-c_p)(\sum_i \min(D_i, C_i) - X'_P)]}{\mathrm{E}[(p-c'_s)((\min(\sum_i D_i, \sum_i C_i) - \sum_i \min(D_i, C_i))]} \\ &= FE(\mathcal{G}(n), c'_s) \end{aligned}$$

The first inequality is because $c_s > c'_s$ and $X_P$ is bounded above by $\sum_i \min(D_i, C_i)$. The second inequality results from the feasibility of $(X_P, X_S)$ to model (3.1) when secondary cost is $c'_s$. ■

Note that when $c_s = c_p$, the expected profit criterion reduces to the expected sales criterion. Hence, by the above proposition, for any fixed flexibility structure $\mathcal{G}(n)$, such that $\mathcal{D}(n) \subseteq \mathcal{G}(n) \subseteq \mathcal{F}(n)$, its flexibility efficiency under the expected profit criterion is never higher than that under the expected sales criterion. This implies that when we take into account the possibility of low response (lower efficiency of secondary production), the value of any flexibility structure may be reduced.[3] This serves as a precaution

[3] Note that the results do not require any assumption about demand or supply, except that the number of products must equal the number of facilities.

not to oversell the benefits of any form of flexibility and a need to examine first the system response level.

Although the previous result may not have come unexpected, a more surprising result is that when system response deteriorates to a certain level (i.e. it enters the low response region), further deterioration will cause no more harm to the system than it does to the full flexibility system. The following theorem captures this insight.

**Theorem 4.** *For a fixed flexibility structure $\mathcal{G}(n)$, such that $\mathcal{D}(n) \subseteq \mathcal{G}(n) \subseteq \mathcal{F}(n)$, $FE(\mathcal{G}(n), c_s)$ is constant over the interval $[\frac{1}{2}(p + c_p), p)$.*

*Proof.* If $c_s \geq \frac{1}{2}(p + c_p)$, then $p - c_p \geq 2(p - c_s)$. This means producing one primary unit is at least as good as producing two secondary units. Therefore, the optimal production allocation can be obtained using a very simple *greedy approach*, that is, the firm must let each facility produce as many units of its primary product as possible, and *only* thereafter use its extra capacity, if any, to produce the extra demand, if any, of its *secondary* product. This means that $X_P = \sum_i \min(D_i, C_i)$ and $X_S$ is also independent of $c_2$. Hence,

$$\begin{aligned} FE(\mathcal{G}(n), c_s) &= \frac{\mathrm{E}[(p - c_s)X_S - (p - c_p)(\sum_i \min(D_i, C_i) - X_P)]}{\mathrm{E}[(p - c_s)(\min(\sum_i D_i, \sum_i C_i) - \sum_i \min(D_i, C_i))]} \\ &= \frac{\mathrm{E}[(p - c_s)X_S]}{\mathrm{E}[(p - c_s)(\min(\sum_i D_i, \sum_i C_i) - \sum_i \min(D_i, C_i))]} \\ &= \frac{\mathrm{E}[X_S]}{\mathrm{E}[\min(\sum_i D_i, \sum_i C_i) - \sum_i \min(D_i, C_i)]} \end{aligned}$$

■

Theorem 4 shows that once system response hits the halfway mark between perfect response and worst-case response, the efficiency of the system plateaus at a certain level. This is because at that point, any additional deterioration in system response can cause only as much harm to the system as it does to the full flexibility system. In the next section, we will revisit this scenario and demonstrate how to obtain this worst-case efficiency level for the chaining structure under special demand distributions.

**Example 2.** *Chaining Strategy for a* $3 \times 3$ *System with Uniform Demand* *Suppose the demands for all products are i.i.d. and uniformly distributed in* $[0, 2\mu]$. *Let* $CE(3, c_s) = FE(\mathcal{C}(3), c_s)$ *be the flexibility efficiency of the 2-chain in a* $3 \times 3$ *system. It is not difficult, though it is tedious, to evaluate* $CE(3, c_s)$ *in closed form for this special case.*

$$CE(3, c_s) = \begin{cases} 1 - \frac{3}{11}\frac{c_s - c_p}{p - c_s} & \text{if } \ c_p \le c_s < \frac{1}{2}(p + c_p) \\ \frac{8}{11} & \text{if } \ \frac{1}{2}(p + c_p) \le c_s < p \end{cases}$$

*Figure 3.1 depicts how the value of* $CE(3, c_s)$ *changes as the cost of secondary production increases or as system response worsens. Note that when* $c_s \approx c_p$, $CE(3, c_s) \approx 100\%$. *Consistent with Theorems 3 and 4,* $CE(3, c_s)$ *is non-increasing in* $c_s$ *over* $[c_p, p)$, *and decreasing over* $[c_p, \frac{1}{2}(p + c_p)]$. *This implies that as the gap in production efficiencies widens (system response deteriorates), the relative value of the chaining strategy decreases. However, this value does not decrease below* $\frac{8}{11}$ *because when the gap reaches a criti-*

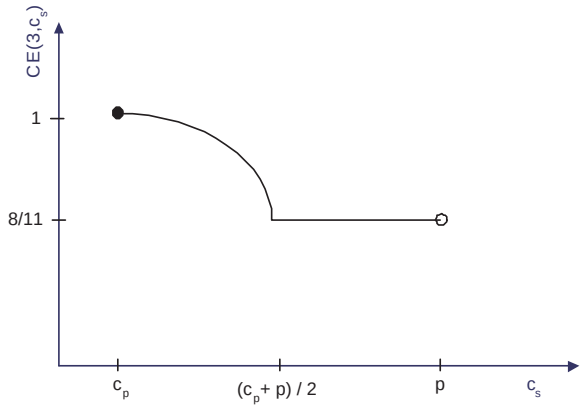


*Fig. 3.1:* Chaining Efficiency vs. Secondary Production Cost ($3 \times 3$ System with Uniform Demand)

*cal level, any further increase in it will have no more effect on the chaining strategy than it would have on full flexibility.*

*This example shows that chaining efficiency may deteriorate as system response worsens. Nonetheless, the performance gap is never worse than 72.7%. However, when n is large, we expect the performance gap to deteriorate further. In the following section, we determine the asymptotic limit of the performance gap when n is large.*

## 3.2 Valuing the Chaining Strategy

In this section, we characterize the asymptotic performance of the chaining strategy for all relevant response levels. For ease of exposition, we consider a stylized model where all products have independent, identically distributed, and symmetric demands, with values in the range $[0, 2\mathrm{E}(D_i)]$. Examples of

such demand distributions are uniform and (truncated) normal distributions. Note that our analysis can be extended easily to more general and asymmetric demand distribution. On the supply side, all facilities have capacities with $C_i = \mathrm{E}(D_i) = \mu,\ \forall\ i$. We call such a system *identical and balanced*.

Because of the assumption of symmetry, we can model demand in the following general form. Let $D_i = \mu + a_i Y_i$, where $0 \leq Y_i \leq \mu$ and

$$a_i = \begin{cases} 1, & \text{with probability } \frac{1}{2} \\ -1, & \text{with probability } \frac{1}{2} \end{cases}$$

Note that $Y_i$ follows some distribution with support $[0, \mu]$ and represents the absolute deviation of demand $D_i$ from the mean $\mu$.

We recall, from Section 1.3.2, the definition of chaining efficiency as follows.

$$CE(n, c_s) = FE(\mathcal{C}(n), c_s) = \frac{\mathrm{E}[\Pi^*_{\mathcal{C}(n)}(\mathbf{D}, c_s)] - \mathrm{E}[\Pi^*_{\mathcal{D}(n)}(\mathbf{D}, c_s)]}{\mathrm{E}[\Pi^*_{\mathcal{F}(n)}(\mathbf{D}, c_s)] - \mathrm{E}[\Pi^*_{\mathcal{D}(n)}(\mathbf{D}, c_s)]}.$$

From (3.3) and (3.4), we can express the denominator of $CE(n, c_s)$ as follows:

$$\mathrm{E}[\Pi^*_{\mathcal{F}(n)}(\mathbf{D}, c_s)] - \mathrm{E}[\Pi^*_{\mathcal{D}(n)}(\mathbf{D}, c_s)] = (p - c_s)\mathrm{E}\left[\min\left(\sum_i D_i, n\mu\right) - \sum_i \min(D_i, \mu)\right] \tag{3.5}$$

The challenge in our analysis is to evaluate the term $\mathrm{E}[\Pi^*_{\mathcal{C}(n)}(\mathbf{D}, c_s)]$ in the numerator. As shown in Theorem 4, the optimal production allocation varies dramatically depending on whether $c_s$ is higher or lower than a threshold value $\frac{1}{2}(p + c_p)$. If $c_s$ is at least the threshold, then $p - c_p \geq 2(p - c_s)$.

This means that it is at least as good to produce one primary unit as it is to produce two secondary units. Therefore, the optimal production allocation can be obtained using a very simple greedy approach. We discuss this further in Section 3.2.1.

On the other hand, if $c_s$ is below the threshold, then $p - c_p < 2(p - c_s)$. This means that it is profitable to replace one unit of primary production with two units of secondary production whenever possible. Unfortunately, obtaining the optimal allocation under this case becomes much more complicated as it depends on the different possible demand realizations. The classic expected sales criterion, whereby $c_s = c_p < \frac{1}{2}(p + c_p)$, is a special instance of the subthreshold case. In Chapter 2, we developed a random walk method to obtain the exact value of $CE(n, c_s)$ for $c_s = c_p$ as $n \to \infty$ for arbitrary demand distributions. Unfortunately, this method does not work for the general subthreshold case when $c_s \neq c_p$. In this chapter, we describe a method to obtain lower bounds for $CE(n, c_s)$ when $c_s \leq \frac{1}{2}(p + c_p)$. To simplify the exposition, we first discuss the lower bound as it applies to the case $c_s = c_p$ in Section 3.2.2. How this method works for the general subthreshold case is presented in Section 3.2.3.

### *3.2.1 System Response is Low*

As mentioned in Theorem 4, the superthreshold case implies that one unit of primary production is at least as profitable as two units of secondary production. Consequently, we use the following greedy approach. The firm must let each facility produce as many units of its primary product as possible, and

*only* thereafter use its extra capacity, if any, to produce the extra demand, if any, of its secondary product. The optimal solution is as follows:

$$\begin{aligned} x^*_{ii} &= \min(D_i, \mu) \quad \forall\ i \\ x^*_{12} &= \min[(D_1 - \mu)^+, (\mu - D_2)^+] \\ x^*_{23} &= \min[(D_2 - \mu)^+, (\mu - D_3)^+] \\ &\vdots \\ x^*_{n-1,n} &= \min[(D_{n-1} - \mu)^+, (\mu - D_n)^+] \\ x^*_{n1} &= \min[(D_n - \mu)^+, (\mu - D_1)^+] \end{aligned} \tag{3.6}$$

The following well-known facts on normally distributed random variables will be useful for our next result.

**Lemma 2.** *If $X, X_1, X_2 \sim N(0, \sigma)$, and $X_1$, $X_2$ are independent, then*

*(a)* $\mathrm{E}[X^+] = \frac{\sigma}{\sqrt{2\pi}}$

*(b)* $\mathrm{E}[\min(X_1^+, X_2^+)] = \frac{\sigma}{\sqrt{2\pi}}(1 - \frac{1}{\sqrt{2}})$

**Theorem 5.** *When $\frac{1}{2}(p + c_p) \leq c_s < p$ and for sufficiently large $n$, the chaining efficiency is decreasing in $n$ and bounded below by the asymptotic chaining efficiency*

$$ACE(c_s) = \lim_{n\to\infty} CE(n, c_s) = \frac{1}{2}\frac{\mathrm{E}[\min(Y_1, Y_2)]}{\mathrm{E}[Y_1]} \leq 0.5.$$

*Proof.* We start by writing the expected optimal chaining profit as follows:

$$\begin{aligned}
\mathrm{E}[\Pi^*_{\mathcal{C}(n)}(\mathbf{D}, c_s)] &= (p-c_p)\mathrm{E}[\textstyle\sum_i \min(D_i, \mu)] \\
&\quad +(p-c_s)\mathrm{E}[\textstyle\sum_i \min[(D_i-\mu)^+, (\mu - D_{i+1})^+]] \\
&= n(p-c_p)\mathrm{E}[\min(D_i, \mu)] \\
&\quad +n(p-c_s)\mathrm{E}[\min[(D_1-\mu)^+, (\mu - D_2)^+]] \\
&= n(p-c_p)\mathrm{E}[\min(D_i, \mu)] + \tfrac{1}{4}n(p-c_s)\mathrm{E}[\min(Y_1, Y_2)]
\end{aligned} \tag{3.7}$$

The first equation is from (3.6), while the second equation comes from the identical distribution of the demands. The last equation is the result of the definition of the absolute demand deviation $Y_i$. Since the first term in (3.7) is also the expected optimal profit for the dedicated system, the numerator of $CE(n, c_s)$ becomes

$$\frac{1}{4}n(p-c_s)\mathrm{E}[\min(Y_1, Y_2)] \tag{3.8}$$

For the denominator, we let $S = \sum_i D_i$. Since $n$ is sufficiently large, we invoke the Central Limit Theorem to get $S \sim N(n\mu, \sqrt{n}\sigma)$ and $X = S - n\mu \sim N(0, \sqrt{n}\sigma)$, where $\sigma$ is the standard deviation of demand $D_i$. Then, we use (3.5), Lemma 2(a), and the definition of $Y_i$ to obtain

$$\begin{aligned}
& (p-c_s)\mathrm{E}[\min(S, n\mu) - \textstyle\sum_i \min(D_i, \mu)] \\
=\ & (p-c_s)\mathrm{E}[\textstyle\sum_i D_i - X^+ - \sum_i D_i + \sum_i (D_i-\mu)^+] \\
=\ & (p-c_s)\left[n\mathrm{E}(D_1-\mu)^+ - \tfrac{\sqrt{n}\sigma}{\sqrt{2\pi}}\right] \\
=\ & n(p-c_s)\left[\tfrac{1}{2}\mathrm{E}[Y_1] - \tfrac{1}{\sqrt{n}}\tfrac{\sigma}{\sqrt{2\pi}}\right]
\end{aligned} \tag{3.9}$$

Combining (3.8) and (3.9) and taking limit, we arrive at the desired result.■

Note that our analysis above does not consider that the chaining strategy usually exploits long chains – chains that link up as many supply and demand nodes as possible. In particular, we only consider that the secondary production arcs in the chaining strategy form a perfect matching. Such a matching can also be formed by a collection of short chains. As a result, a long chain performs identically to short chains under this scenario. This observation contrasts with the fundamental insight in the literature when we assume that system response is perfect. See Figure 3.2 for an illustration.

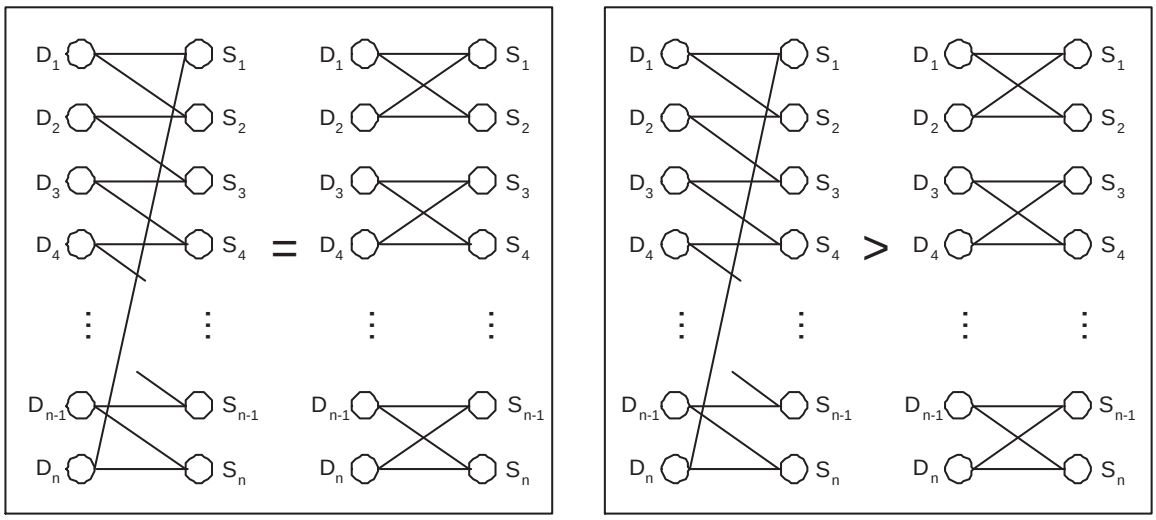


*Fig. 3.2:* Long Chain vs. Short Chains: The Effect of System Response

### 3.2.2 System Response is Perfect

When $c_s = c_p$, the expected profit criterion is equivalent to the expected sales criterion. Hence, expected optimal profit becomes expected maximum flow. We let $MF(\mathcal{D}(n), \mathbf{D})$, $MF(\mathcal{C}(n), \mathbf{D})$, and $MF(\mathcal{F}(n), \mathbf{D})$ be the *maximum dedicated flow*, the *maximum chaining flow*, and the *maximum full flexibility flow*, respectively, for an $n \times n$ system with demand realization $\mathbf{D}$. It follows that the *chaining efficiency* can be written as

$$CE(n, c_p) = \frac{\mathrm{E}[MF(\mathcal{C}(n), \mathbf{D})] - \mathrm{E}[MF(\mathcal{D}(n), \mathbf{D})]}{\mathrm{E}[MF(\mathcal{F}(n), \mathbf{D})] - \mathrm{E}[MF(\mathcal{D}(n), \mathbf{D})]}$$

Similar to (3.9), we can express the denominator as

$$\mathrm{E}[MF(\mathcal{F}(n), \mathbf{D})] - \mathrm{E}[MF(\mathcal{D}(n), \mathbf{D})] = n\left[\frac{1}{2}\mathrm{E}[Y_1] - O(1/\sqrt{n})\right] \qquad (3.10)$$

Let

$$\text{Expected Chaining Gain} \;=\; \mathrm{E}[MF(\mathcal{C}(n), \mathbf{D}) - MF(\mathcal{D}(n), \mathbf{D})]$$

Consider any demand realization $\mathbf{D}$. Observe that each demand node $i$ has either $D_i > \mu$ (positive node) or $D_i < \mu$ (negative node) with equal likelihood.[4] We define a *cluster* to be a run of consecutive positive nodes followed by a run of consecutive negative nodes. For example, suppose $n = 10$ and the demand outcome is $\{N, P, P, P, N, N, N, P, N, N\}$ where $P$ denotes

[4] We assume demand distribution has continuous support.

a positive node while $N$ denotes a negative node. The 2nd to 7th nodes form the first cluster $\{P, P, P, N, N, N\}$ while the last 3 and the 1st form the next cluster $\{P, N, N, N\}$. This allows us to break the whole system into smaller pieces (called clusters), and we can easily optimize the flow for each cluster. The aggregate solution from all clusters remains feasible for the max-flow problem of the whole system, and thus provides a lower bound for $MF(\mathcal{C}(n), \mathbf{D})$, that is,

$$\begin{aligned} \text{Expected Chaining Gain} &\geq \text{E[Sum of Cluster Chaining Gains]} \\ &= \text{E[Number of Clusters]} \cdot \text{E[Cluster Chaining Gain]} \end{aligned} \tag{3.11}$$

The last equation holds for large $n$, and is the result of Wald's equation (see Ross [46], pg 462) and the fact that all clusters are probabilistically identical and independent.

To obtain the expected cluster chaining gain, we consider just one cluster. Observe that for this cluster, the lengths of the positive and negative runs as well as the deviations of realized demands from the mean are all random variables. We let $M$ and $N$ be the lengths of the positive and negative runs, respectively. Both $M$ and $N$ follow a geometric distribution with $p = 0.5$. From our earlier definition, we have $Y_i$ for the demand deviations ($Y_i = D_i - \mu$ for positive nodes while $Y_i = \mu - D_i$ for negative nodes).

**Lemma 3.** *For any cluster with $M$ positive nodes followed by $N$ negative nodes, the maximum chaining flow is*

$$MF(\mathcal{C}(n), \mathbf{D}) = \min\left(\sum_{i=1}^{M+N} D_i, (M+1)\mu + \sum_{i=M+1}^{M+N} D_i, (M+N)\mu\right)$$

*and the cluster chaining gain is*

$$\textit{Cluster Chaining Gain} = \min\left(\sum_{i=1}^{M} Y_i, \mu, \sum_{i=M+1}^{M+N} Y_i\right)$$

*Proof.* We use the equivalence between max-flow and min-cut to derive the above result. Consider the max-flow problem on the network shown in Figure 3.3, from source $s$ to sink $t$. All links are directed from left to right with capacities as indicated. Arcs from demand nodes to supply nodes have infinite capacities.

Let $\mathbb{C}$ be a cut of the network and $V(\mathbb{C})$ be its cut value. Because of the infinite capacities of demand-to-supply arcs, every cut with finite cut value can be uniquely represented by $\{s\}$ union with a subset of $\mathbb{S} = \{1, 2, \ldots, M+N\}$. For example, $\mathbb{C}_1 = \{s, 1, 2, \ldots, M-1, M+N\}$ represents the cut in Figure 3.3, with cut value $V(\mathbb{C}_1) = \sum_{i=M}^{M+N-1} D_i + (M+1)\mu$.

Let $\mathbb{S}^+ = \{1, 2, \ldots, M\}$ and $\mathbb{S}^- = \{M+1, M+2, \ldots, M+N\}$. Then every cut can be written as $\mathbb{C} = \{s\} \cup \mathbb{P} \cup \mathbb{N}$ where $\mathbb{P} \subseteq \mathbb{S}^+, \mathbb{N} \subseteq \mathbb{S}^-$. Recall that $D_i > \mu$ for $i \in \mathbb{S}^+$ and $D_i < \mu$ for $i \in \mathbb{S}^-$. We show next that essentially we only need

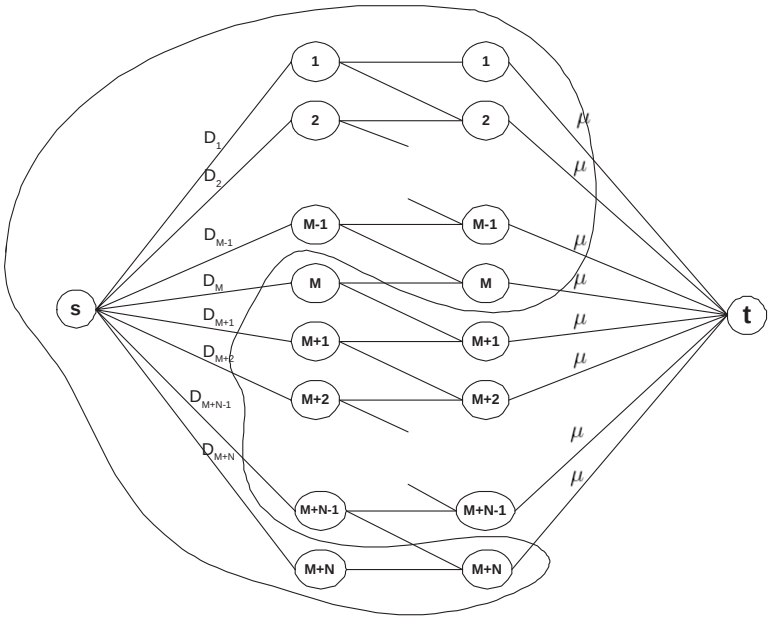


*Fig. 3.3:* Sample Cut for Network with Perfect System Response: $\mathbb{C}_1 = \{s, 1, 2, \ldots, M-1, M+N\}$

to keep track of the values of three different cuts. Let $\mathbb{C}^* = \{s\} \cup \mathbb{P}^* \cup \mathbb{N}^*$ be a minimum cut of the above network flow problem.

- If $\mathbb{P}^* = \emptyset$, then $\mathbb{N}^* = \emptyset$, thus $\mathbb{C}^* = \{s\}$.

  Suppose $\mathbb{N}^* \neq \emptyset$. Identify $j = \min\{i : i \in \mathbb{N}^*\}$. Let $\mathbb{C} = \{s\} \cup \mathbb{N}$ where $\mathbb{N} = \mathbb{N}^* \backslash \{j\}$. Then $V(\mathbb{C}) = V(\mathbb{C}^*) + D_j - \mu$ if $j+1 \in \mathbb{N}^*$, while $V(\mathbb{C}) = V(\mathbb{C}^*) + D_j - 2\mu$ if $j+1 \notin \mathbb{N}^*$. Hence, $V(\mathbb{C}) < V(\mathbb{C}^*)$, which contradicts the optimality of $\mathbb{C}^*$.

- If $\mathbb{P}^* \neq \emptyset$, then $\mathbb{P}^* = \mathbb{S}^+$.

  Suppose $\emptyset \subset \mathbb{P}^* \subset \mathbb{S}^+$. If $1 \in \mathbb{P}^*$, then let $j = \min\{i : i \notin \mathbb{P}^*\}$ and $\mathbb{C} = \mathbb{C}^* \cup \{j\}$. It follows that $V(\mathbb{C}) = V(\mathbb{C}^*) - D_j$ if $j+1 \in \mathbb{C}^*$, while $V(\mathbb{C}) = V(\mathbb{C}^*) - D_j + \mu$ if $j+1 \notin \mathbb{C}^*$. Hence, $V(\mathbb{C}) < V(\mathbb{C}^*)$, which

contradicts the optimality of $\mathbb{C}^*$.

If $1 \notin \mathbb{P}^*$, then let $j = \min\{i : i \in \mathbb{P}^*\} - 1$ and $\mathbb{C} = \mathbb{C}^* \cup \{j\}$. It follows that $V(\mathbb{C}) = V(\mathbb{C}^*) - D_j$ if $j - 1 \in \mathbb{C}^*$, while $V(\mathbb{C}) = V(\mathbb{C}^*) - D_j + \mu$ if $j - 1 \notin \mathbb{C}^*$. Hence, $V(\mathbb{C}) < V(\mathbb{C}^*)$, which contradicts the optimality of $\mathbb{C}^*$.

- If $\mathbb{N}^* \neq \emptyset$, then $\mathbb{P}^* = \mathbb{S}^+$ and $\mathbb{N}^* = \mathbb{S}^-$, thus $\mathbb{C}^* = \{s\} \cup \mathbb{S}$.

  $\mathbb{P}^* = \mathbb{S}^+$ follows from the previous two statements. Therefore, it suffices to show that $\mathbb{N}^* \neq \emptyset$ and $\mathbb{P}^* = \mathbb{S}^+$ implies $\mathbb{N}^* = \mathbb{S}^-$. Suppose $\emptyset \subset \mathbb{N}^* \subset \mathbb{S}^-$. If $M + N \in \mathbb{N}^*$, then let $j = \max\{i : i \notin \mathbb{N}^*\} + 1$ and $\mathbb{C} = \mathbb{C}^* \backslash \{j\}$. It follows that $V(\mathbb{C}) = V(\mathbb{C}^*) + D_j - \mu < V(\mathbb{C}^*)$, which contradicts the optimality of $\mathbb{C}^*$.

  If $M+N \notin \mathbb{N}^*$, then let $j = \max\{i : i \in \mathbb{N}^*\}$ and $\mathbb{C} = \mathbb{C}^* \backslash \{j\}$. It follows that $V(\mathbb{C}) = V(\mathbb{C}^*) + D_j - \mu$ if $j - 1 \in \mathbb{C}^*$, while $V(\mathbb{C}) = V(\mathbb{C}^*) + D_j - 2\mu$ if $j - 1 \notin \mathbb{C}^*$. Hence, $V(\mathbb{C}) < V(\mathbb{C}^*)$, which contradicts the optimality of $\mathbb{C}^*$.

  Note that if $\mathbb{N}^* = \mathbb{S}^-$, then deleting any node $j \in \mathbb{N}^*$ from $\mathbb{C}^*$ does not produce a cut with lower value. Thus, the argument holds because of $\emptyset \subset \mathbb{N}^* \subset \mathbb{S}^-$.

Hence, the search for a minimum cut can be restricted to the three cuts $\{s\}$, $\{s\} \cup \mathbb{S}^+$, and $\{s\} \cup \mathbb{S}$, from which the first result follows.

The second result is an easy consequence of the first, since

$$\begin{aligned}
\text{Cluster Chaining Gain} &= MF(\mathcal{C}(n), \mathbf{D}) - MF(\mathcal{D}(n), \mathbf{D}) \\
&= \min\left( \sum_{i=1}^{M+N} D_i, (M+1)\mu + \sum_{i=M+1}^{M+N} D_i, (M+N)\mu \right) \\
&\quad - \left( \sum_{i=1}^{M} \mu + \sum_{i=M+1}^{M+N} D_i \right) \\
&= \min\left( \sum_{i=1}^{M} Y_i, \mu, \sum_{i=M+1}^{M+N} Y_i \right)
\end{aligned}$$

■

**Lemma 4.** *For an $n \times n$ system,*

$$\frac{E[\textit{Number of Clusters}]}{n} \to \frac{1}{4} \quad \textit{as } n \to \infty$$

*Proof.* Without loss of generality, assume node 1 is negative. Then, the number of clusters from node 1 to node $n$ can be viewed as a counting process, in fact, a renewal process whereby each occurrence of a cluster constitutes a renewal. By the Elementary Renewal Theorem (see Ross [46], pg 409) and because E[cluster length] = E$[M + N] = 4$, the result follows. ■

Combining (3.10), (3.11), Lemma 3, and Lemma 4, we obtain the following key result of the paper.

**Theorem 6.** *When the cost of secondary production equals the cost of primary production ($c_s = c_p$), the asymptotic chaining efficiency is bounded*

*below as follows:*

$$ACE(c_p) \geq \frac{1}{2}\frac{\mathrm{E}[\min(\sum_{i=1}^{M} Y_i, \sum_{i=1}^{N} \tilde{Y}_i, \mu)]}{\mathrm{E}[Y_1]}$$

*where* $M, N$ *are geometric r.v. with* $p = 0.5$*, and* $Y_i$*,* $\tilde{Y}_i$ *are i.i.d. random variables with support* $[0, \mu]$*.*

### *3.2.3 System Response is High*

As mentioned earlier, the subthreshold case means that it is profitable to displace one unit of primary production in favor of two units of secondary production. This implies that the greedy approach used in Section 3.2.1 no longer works. The maximum flow approach used in Section 3.2.2 must be adjusted to apply to this case because maximum flow includes using the extra capacity of facility $i$ to meet the extra demand for product $i+j$ for any $i$ and $j$. We call such an allocation a *$j$-order displacement*. It is easy to see that a $j$-order displacement is justified only if $j$ units of secondary production are as profitable as $j-1$ units of primary production, a requirement not necessarily satisfied by the subthreshold condition. Consequently, our analysis requires dividing this subthreshold case into countably infinite subcases, namely,

$$\frac{k}{k-1} \leq \frac{p-c_p}{p-c_s} < \frac{k-1}{k-2} \tag{3.12}$$

for $k = 3, 4, \ldots$

For subcase $k$, a $j$-order displacement is profitable for $j < k$, but not

for $j \geq k$. Therefore, if we use the maximum flow approach, we should distinguish flows that result in profitable displacement from flows that do not. In particular, the optimal allocation should not include displacements of order $k$ or higher. Secondly, the flows must be assigned different weights, corresponding to different profit levels, depending on the amount of production displaced. Specifically, for a $j$-order displacement, the unit profit is $j(p - c_s) - (j-1)(p - c_p)$.

Let $g(j)$ be the expected maximum flow net of dedicated flow for a cluster of length $j$. We obtain $g(j)$ in a manner similar to Section 3.2.2 and set $g(1) = 0$ for completeness. Therefore, $\Delta g(j) = g(j+1) - g(j)$ represents the incremental flow in a cluster of length $j+1$ over a cluster of length $j$. This additional flow is made up of displacements of order $j$ or less, which implies a unit profit level of at least $j(p-c_s)-(j-1)(p-c_p)$. However, this is relevant only for $j < k$. For $j \geq k$, the incremental flow may be zero if all displacements are of order $j$, and thus unprofitable. Recalling from Section 3.2.2 that $M + N$ denotes cluster length and writing Cluster Chaining Gain as CCG, we have the following recursive inequalities:
For $j = 1, 2, \ldots, k-1$,

$$\begin{aligned} \mathrm{E}[\mathrm{CCG} \mid M+N = j+1] \quad \geq \quad & \mathrm{E}[\mathrm{CCG} \mid M+N = j] \\ & +[j(p-c_s) - (j-1)(p-c_p)]\Delta g(j) \end{aligned} \tag{3.13}$$

For $j = k, k+1, \ldots$

$$\mathrm{E}[\mathrm{CCG} \mid M+N = j+1] \geq \mathrm{E}[\mathrm{CCG} \mid M+N = j] \tag{3.14}$$

where $\mathrm{E}[\mathrm{CCG} \mid M+N=1]=0$. It follows from (3.13) and (3.14) that for $j=1,2,\ldots,k-1$,

$$\mathrm{E}[\mathrm{CCG} \mid M+N=j] \geq \textstyle\sum_{i=1}^{j-1}[i(p-c_s)-(i-1)(p-c_p)]\Delta g(i) \qquad (3.15)$$

while for $j=k,k+1,\ldots$

$$\mathrm{E}[\mathrm{CCG} \mid M+N=j] \geq \textstyle\sum_{i=1}^{k-1}[i(p-c_s)-(i-1)(p-c_p)]\Delta g(i) \qquad (3.16)$$

At this point, we are ready to prove the following theorem.

**Theorem 7.** *For $k=3,4,\ldots$, if $\frac{1}{k}[p+(k-1)c_p] \leq c_s < \frac{1}{k-1}[p+(k-2)c_p]$, then*

$$ACE(c_s) \geq \frac{\sum_{i=1}^{k-1}[i(p-c_s)-(i-1)(p-c_p)]\Delta g(i)\cdot \mathrm{P}\{M+N>i\}}{2(p-c_s)\mathrm{E}[Y_1]}$$

*where*

$\Delta g(j)=g(j+1)-g(j)$, $g(j)=\mathrm{E}[\min(\sum_{i=1}^{M}Y_i,\sum_{i=1}^{N}\tilde{Y}_i,\mu) \mid M+N=j]$, $g(1)=0$, *and $M,N$ are geometric with $p=0.5$, and $Y_i$, $\tilde{Y}_i$ are i.i.d. random variables with support $[0,\mu]$.*

*Proof.* Conditioning on $M+N$ and using (3.15) and (3.16), we get

$$\begin{aligned}
&\text{E[Cluster Chaining Gain]} \\
=\ &\sum_{j=2}^{k} \text{E[CCG} \mid M+N=j] \cdot \text{P}\{M+N=j\} \\
&+ \sum_{j=k+1}^{\infty} \text{E[CCG} \mid M+N=j] \cdot \text{P}\{M+N=j\} \\
\geq\ &\sum_{j=2}^{k}\sum_{i=1}^{j-1}[i(p-c_s)-(i-1)(p-c_p)]\Delta g(i) \cdot \text{P}\{M+N=j\} \\
&+\sum_{i=1}^{k-1}[i(p-c_s)-(i-1)(p-c_p)]\Delta g(i) \cdot \text{P}\{M+N>k\} \\
=\ &\sum_{i=1}^{k-1}[i(p-c_s)-(i-1)(p-c_p)]\Delta g(i) \cdot \text{P}\{i<M+N\leq k\} \\
&+\sum_{i=1}^{k-1}[i(p-c_s)-(i-1)(p-c_p)]\Delta g(i) \cdot \text{P}\{M+N>k\} \\
=\ &\sum_{i=1}^{k-1}[i(p-c_s)-(i-1)(p-c_p)]\Delta g(i) \cdot \text{P}\{M+N>i\} \qquad (3.17)
\end{aligned}$$

Next, the given condition implies (3.12), which means that only displacements of an order less than $k$ are profitable. Following arguments similar to Lemma 3, Lemma 4, and Theorem 6, we obtain

$$\lim_{n\to\infty} CE(n, c_s) \geq \frac{\text{E[Cluster Chaining Gain]}}{2(p-c_s)E[Y_1]} \qquad (3.18)$$

Substituting (3.17) into (3.18), we arrive at the desired result. ■

Theorems 5, 6, and 7 complete the characterization of asymptotic chaining efficiency over all relevant response levels. In the next section, we demonstrate how these results can be applied to two commonly used distributions, namely, uniform and normal.

### 3.2.4 Computational Examples

Applying Theorem 5 to the uniform and the normal distributions, we obtain the following expressions:

1. For uniform distribution, we have $Y_i \sim U(0, \mu)\ \forall\ i$ as well as

$$\begin{aligned}
\mathrm{E}[Y_1] &= \int_0^{\mu} y \cdot \frac{1}{\mu} dy \;=\; \frac{1}{2}\mu \\
\mathrm{E}[\min(Y_1, Y_2)] &= 2\int_0^{\mu}\int_0^{y_1} y_2 \cdot \frac{1}{\mu} \cdot \frac{1}{\mu} dy_2 dy_1 \;=\; \frac{1}{3}\mu \\
\mathrm{E}[D_i^2] &= \int_0^{2\mu} y^2 \cdot \frac{1}{2\mu} dy \;=\; \frac{4}{3}\mu^2 \\
\sigma &= \sqrt{\mathrm{E}[D_i^2] - \mathrm{E}[D_i]^2} \;=\; \sqrt{\frac{4}{3}\mu^2 - \mu^2} \;=\; \frac{\mu}{\sqrt{3}}
\end{aligned}$$

Therefore, for $\frac{1}{2}(p + c_p) \leq c_s < p$

$$CE(n, c_s) = \frac{\frac{1}{4}\mathrm{E}[\min(Y_1, Y_2)]}{\frac{1}{2}\mathrm{E}[Y_1] - \frac{1}{\sqrt{n}}\frac{\sigma}{\sqrt{2\pi}}} = \frac{1}{3 - \frac{2\sqrt{6}}{\sqrt{n}\sqrt{\pi}}}$$

and

$$ACE(c_s) = \frac{1}{3} \approx 33.33\%.$$

2. For normal distribution, we have $D_i \sim N(\mu, \sigma)\ \forall\ i$. It follows that $X_i = D_i - \mu \sim N(0, \sigma)$ and $Y_i = |X_i|$, $\forall\ i$. Assume $\mu \geq 3\sigma$ such that negative demand has negligible probability.

$$\begin{aligned}
\mathrm{E}[Y_1] &= \mathrm{E}[X_1^+] + \mathrm{E}[X_1^-] = 2\mathrm{E}[X_1^+] = \frac{2\sigma}{\sqrt{2\pi}} \\
\mathrm{E}[\min(Y_1, Y_2)] &= \mathrm{E}[\min(X_1^+, X_2^+)] + \mathrm{E}[\min(X_1^+, X_2^-)] \\
&\quad + \mathrm{E}[\min(X_1^-, X_2^+)] + \mathrm{E}[\min(X_1^-, X_2^-)] \\
&= 4\mathrm{E}[\min(X_1^+, X_2^+)] = \frac{4\sigma}{\sqrt{2\pi}}\left(1 - \frac{1}{\sqrt{2}}\right)
\end{aligned}$$

Therefore, $\frac{1}{2}(p + c_p) \leq c_s < p$

$$CE(n, c_s) = \frac{\frac{1}{4}\mathrm{E}[\min(Y_1, Y_2)]}{\frac{1}{2}\mathrm{E}[Y_1] - \frac{1}{\sqrt{n}}\frac{\sigma}{\sqrt{2\pi}}} = \frac{1 - \frac{1}{\sqrt{2}}}{1 - \frac{1}{\sqrt{n}}}$$

and

$$ACE(c_s) = 1 - \frac{1}{\sqrt{2}} \approx 29.29\%$$

We can then apply Theorems 6 and 7 via Monte Carlo sampling. The following table summarizes the computational results.

| Theorem | Response | Uniform Distribution | Normal Distribution |
|---|---|---|---|
| Theorem 5 | Low | 33.33% | 29.29% |
| Theorem 6 | Perfect | 52.70% | 53.85% |
| Theorem 7 | High | See Figure 3.4 | See Figure 3.4 |

*Tab. 3.2:* Asymptotic Chaining Efficiency for all Relevant System Response Levels (Uniform and Normal Demands)

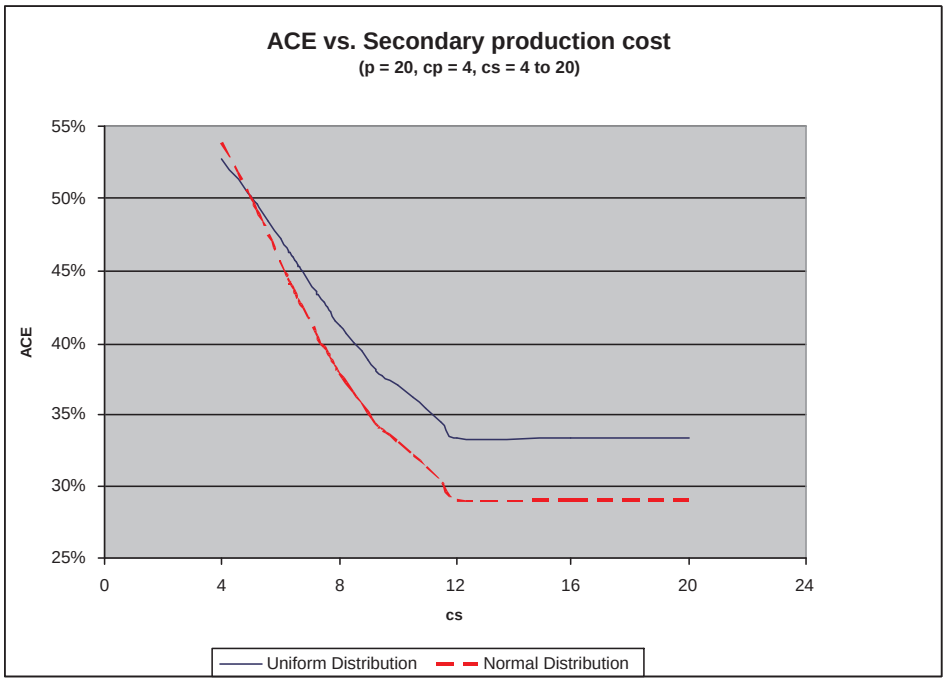


*Fig. 3.4:* Bounds for Asymptotic Chaining Efficiency vs. Secondary Production Cost (Uniform and Normal Demands)

## 3.3 Trade-offs and Complements

### 3.3.1 Range versus Response

In Theorem 5 and Section 3.2.4, we have already seen that a system with low range and low response (chaining with high $c_s$) can perform quite badly (e.g., 29.29% for normally distributed demands). To improve such a system, one can either upgrade response or upgrade range. With limited resources, it is of interest to know which upgrade provides greater improvement: a high response with limited range or a high range with low response, that is, chaining with low secondary cost or full flexibility with high secondary cost.

Let $\mathcal{S}_1(n)$ and $\mathcal{S}_2(n)$ be the high response (chaining) and high range (full flexibility) systems, respectively. Denote their respective costs of secondary production by $c_1$ and $c_2$ such that $c_1 < c_2$. Our goal then is to compare the ratios of each system to the best possible system, which is full flexibility with secondary cost at $c_p$. That is,

$$\lim_{n\to\infty} \frac{\mathrm{E}[\Pi^*_{\mathcal{S}_1(n)}(\mathbf{D}, c_1)]}{\mathrm{E}[\Pi^*_{\mathcal{F}(n)}\mathbf{D}, c_p)]} \quad \text{versus} \quad \lim_{n\to\infty} \frac{\mathrm{E}[\Pi^*_{\mathcal{S}_2(n)}(\mathbf{D}, c_2)]}{\mathrm{E}[\Pi^*_{\mathcal{F}(n)}(\mathbf{D}, c_p)]}$$

Suppose further that $c_2 \geq \frac{1}{2}(p + c_p)$ and $c_1 = c_p$. It is easy to see that

$$\begin{aligned}
\lim_{n\to\infty} \frac{\mathrm{E}[\Pi^*_{\mathcal{S}_1(n)}(\mathbf{D}, c_1)]}{\mathrm{E}[\Pi^*_{\mathcal{F}(n)}\mathbf{D}, c_p)]} &= ASR(\mathcal{C}(\infty)) \\
&= \mathrm{ACE} + (1 - \mathrm{ACE})\left(\frac{\mu - E[(\mu - D_i)^+]}{\mu}\right) \\
&= 1 - (1 - \mathrm{ACE}) \cdot \frac{E[(\mu - D_i)^+]}{\mu} \qquad (3.19)
\end{aligned}$$

Moreover, we can prove a bound on $ACE$ following the random walk approach introduced in Chapter 2.

**Lemma 5.**

$$ACE = 1 - \frac{E[\psi_0]}{2E[\tau_0]\mathrm{E}[(D_i - \mu)^+]} \geq \frac{1}{2}.$$

*where $\psi_0$ and $\tau_0$ are the cycle overshoot and cycle duration in the random walk approach used in Chapter 2.*

*Proof.* The result follows from Wald's identity, the symmetry of demand distribution, and

$$\psi_0 \leq \sum_{i=1}^{\tau_0}(D_i - \mu)^- + (D_{\tau_0} - \mu)^+ = \sum_{i=1}^{\tau_0 - 1}(D_i - \mu)^- + (D_{\tau_0} - \mu)^+$$

■

We are now ready to present the following result:

**Theorem 8.** *If demands are i.i.d. and symmetric, then response is at least as good as range, that is,*

$$\lim_{n\to\infty} \frac{\mathrm{E}[\Pi^*_{\mathcal{S}_1(n)}(\mathbf{D}, c_1)]}{\mathrm{E}[\Pi^*_{\mathcal{F}(n)}\mathbf{D}, c_p)]} \geq \lim_{n\to\infty} \frac{\mathrm{E}[\Pi^*_{\mathcal{S}_2(n)}(\mathbf{D}, c_2)]}{\mathrm{E}[\Pi^*_{\mathcal{F}(n)}(\mathbf{D}, c_p)]}$$

*where response is chaining with secondary cost at $c_1 = c_p$ while range is full flexibility with secondary cost at $c_2 \geq \frac{1}{2}(p + c_p)$.*

*Proof.* Using Lemma 5, equation (3.19), and $c_2 \geq \frac{1}{2}(p + c_p)$,

$$
\begin{aligned}
& \lim_{n\to\infty} \frac{\mathrm{E}[\Pi^*_{\mathcal{S}_2(n)}(\mathbf{D}, c_2)]}{\mathrm{E}[\Pi^*_{\mathcal{F}(n)}(\mathbf{D}, c_p)]} \\
= & \lim_{n\to\infty} \frac{(p-c_p)\mathrm{E}[\sum_{i=1}^n \min(D_i,\mu)] + (p-c_2)\mathrm{E}[\min(\sum_{i=1}^n D_i, n\mu) - \sum_{i=1}^n \min(D_i,\mu)]}{(p-c_p)\mathrm{E}[\min(\sum_{i=1}^n D_i, n\mu)]} \\
\leq & \lim_{n\to\infty} \frac{(p-c_p)\mathrm{E}[\sum_{i=1}^n \min(D_i,\mu)] + \frac{1}{2}(p-c_p)\mathrm{E}[\min(\sum_{i=1}^n D_i, n\mu) - \sum_{i=1}^n \min(D_i,\mu)]}{(p-c_p)\mathrm{E}[\min(\sum_{i=1}^n D_i, n\mu)]} \\
= & \lim_{n\to\infty} \frac{\frac{1}{2}\mathrm{E}[\min(\sum_{i=1}^n D_i, n\mu) + \frac{1}{2}\mathrm{E}[\sum_{i=1}^n \min(D_i,\mu)]}{\mathrm{E}[\min(\sum_{i=1}^n D_i, n\mu)]} \\
= & \frac{1}{2} + \frac{1}{2}\lim_{n\to\infty} \frac{\mathrm{E}[\sum_{i=1}^n \min(D_i,\mu)]}{\mathrm{E}[\min(\sum_{i=1}^n D_i, n\mu)]} \\
= & 1 - \frac{1}{2}\cdot\frac{E[(\mu - D_i)^+]}{\mu} \\
\leq & 1 - (1 - \mathrm{ACE})\cdot\frac{E[(\mu - D_i)^+]}{\mu} \\
= & \lim_{n\to\infty} \frac{\mathrm{E}[\Pi^*_{\mathcal{S}_1(n)}(\mathbf{D}, c_1)]}{\mathrm{E}[\Pi^*_{\mathcal{F}(n)}\mathbf{D}, c_p)]}
\end{aligned}
$$

■

**Remark 1.** *If the response of the high range system improves, that is, $c_2 = \frac{1}{2}(p + c_p) - \epsilon$, and $ACE = \frac{1}{2}$ (e.g., 2-point distribution), then Theorem 8 no longer holds.*

**Remark 2.** *In general, when $c_2 = \frac{p+ac_p}{a+1}$ for $a > 1$, Theorem 8 no longer holds for an arbitrary demand distribution (i.e., arbitrary ACE). Nevertheless, Theorem 8 still holds if $ACE \geq \frac{a}{a+1}$. This can be achieved by reducing the demand coefficient of variation. In other words, reducing demand vari-*

*ability allows the high response system to outperform an even more responsive high range system.*

**The Impact of Size, Demand Variability, and Cost**

The above theorem compares asymptotic performances, that is, when $n \to \infty$. For $n \to \infty$, Theorem 8 tells us that chaining with perfect response ($c_1 = c_p$) can beat full flexibility with secondary cost at $c_2 \geq \frac{1}{2}(p + c_p)$, for any symmetric distribution. To illustrate, suppose $p = 10$ and $c_p = 5$. Then to achieve the same performance as chaining with perfect response, one must install full flexibility and yet can let response slip only halfway (i.e. secondary cost can increase up to 7.5 only).

Naturally, one may wonder whether this result still holds when system size is finite. More precisely, when $n$ is finite, how much room in response are we allowed to slip if we replace chaining with perfect response by full flexibility? In other words, we seek the least secondary cost $c_2^*$ of a full flexibility system that chaining with perfect response can still beat. Without loss of generality, we let $p = 1$ and $c_p = 0$. Then, Theorem 8 states that $c_2^* = 0.5$. Figures 3.5 and 3.6 show the values of $c_2^*$ as $n$ range from 10 to 40 for discrete uniform and normal demands, respectively.

The figures suggest that as system size decreases, the value of upgrading to full flexibility also decreases. To put in perspective, take the example of $n = 20$, $p = 10$, and $c_p = 5$. If demand follows a 10-point distribution, then upgrading to full flexibility allows the firm to relax system response from 5 to 6.5. This is a paltry gain compared to the ridiculous expense involved in installing 400 production links. This implies that Theorem 8 already

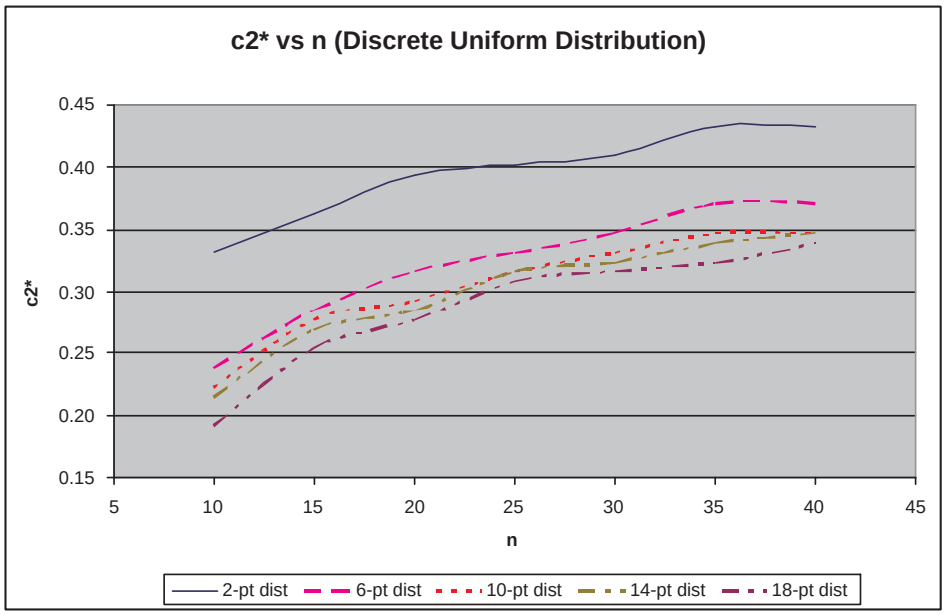


*Fig. 3.5:* Full Flexibility's Least Secondary Production Cost vs. System Size (Discrete Uniform Demand)

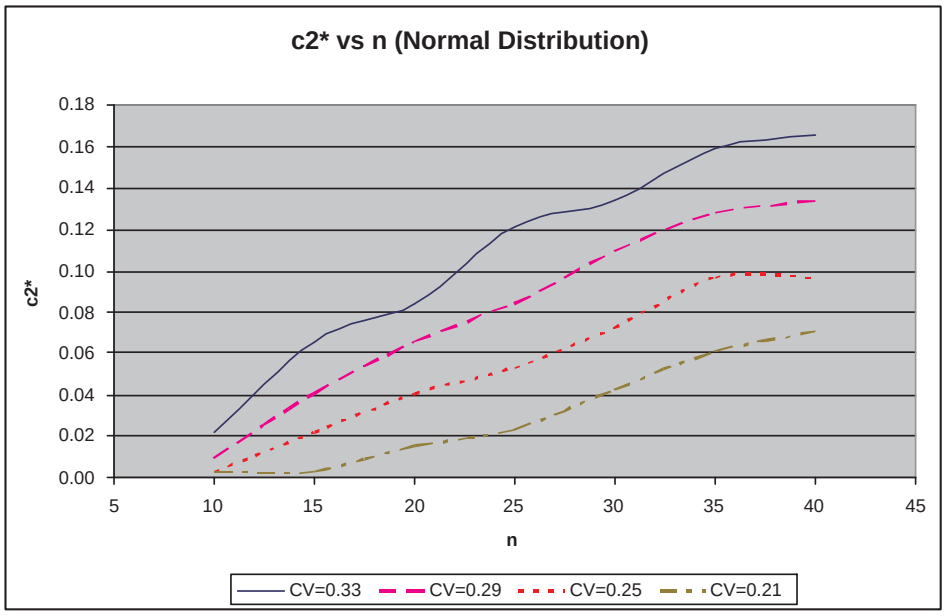


*Fig. 3.6:* Full Flexibility's Least Secondary Production Cost vs. System Size (Normal Demand)

considers the worst case, and thus offers a conservative conclusion. Moreover, as the demand coefficient of variation[5] decreases, the value of upgrading to full flexibility likewise decreases. This suggests the complementary nature of upgrading response and reducing demand variability. We will have more to say on this in Section 3.3.2.

However, it is possible that perfect response is not achievable. If this happens, we expect $c_2^*$ to rise as the chaining system departs from perfect response. In other words, it becomes harder for chaining to beat full flexibility. Figures 3.7 and 3.8 illustrate this pattern. In addition, the gradient of the curve also provides some insight into choosing a system when perfect response is not possible. It is easy to argue that the higher the gradient, the more unstable the chaining system. For example, if the gradient is significantly greater than 1, then deterioration of the chaining system response by a little makes full flexibility significantly more attractive. Now, we see from the figures that the gradient value increases as the demand coefficient of variation (CV) decreases. Moreover, if we expect system response to deteriorate over time (secondary cost to go up), then we should go for the more stable system. Otherwise, we should prefer the less stable system. These observations lead to the following guidelines listed in Table 3.3.

| | Response will deteriorate | Response will improve |
|---|---|---|
| CV high | Prefer chaining | Prefer full flexibility |
| CV low | Prefer full flexibility | Prefer chaining |

*Tab. 3.3:* System Choice without Perfect Response

[5] The series of discrete uniform distributions in Figure 3.5 are arranged in decreasing coefficient of variation.

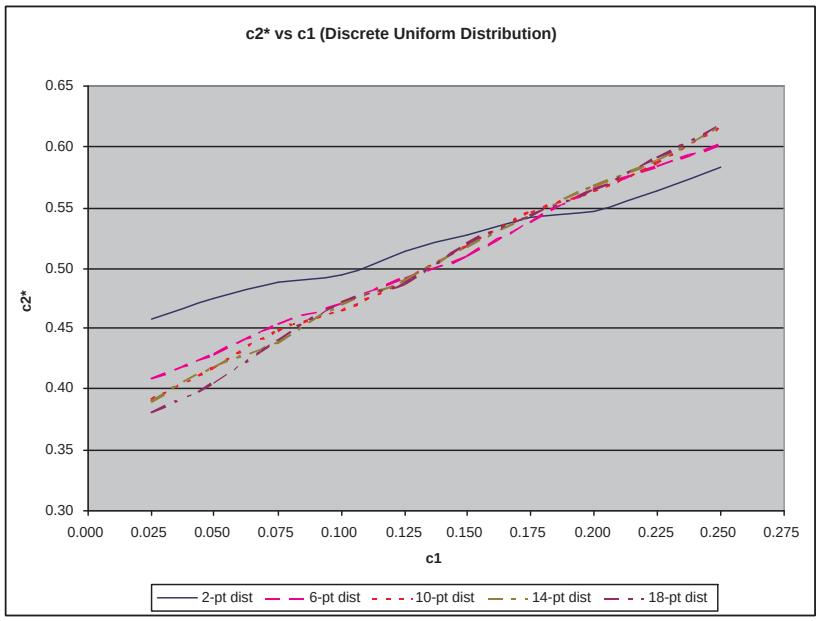


*Fig. 3.7:* Full Flexibility's Least Secondary Production Cost vs. Partial Flexibility's Secondary Production Cost (Discrete Uniform Demand)

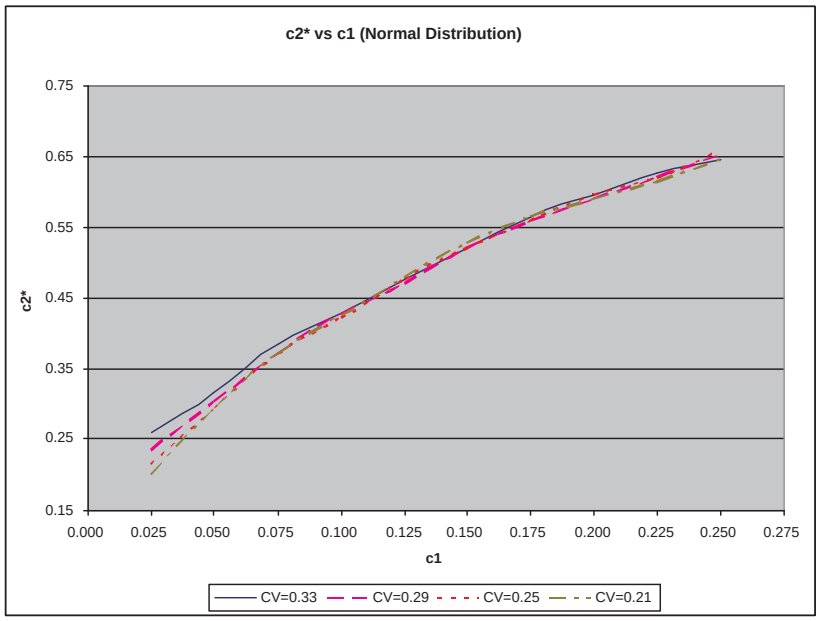


*Fig. 3.8:* Full Flexibility's Least Secondary Production Cost vs. Partial Flexibility's Secondary Production Cost (Normal Demand)

**Asymmetry and Correlation**

Obviously, one would next wonder whether these results carry over beyond the identical and balanced case. We address this concern by studying the 16-product 8-plant example used by Jordan and Graves [32]. Figure 3.9 shows the plants and products used in this setting. It also includes expected demand for each product, capacity for each plant, as well as both primary (dotted lines) and secondary (full lines) production links. As in [32], we also assume that product demands are truncated ($\pm 2\sigma$) normally distributed random variables with standard deviation $\sigma_i = 0.4\mathrm{E}[D_i]$. The products can be divided into three subgroups: products A to F, products G to M, and products N to P. The demand for products in the same subgroup are pairwise correlated with a correlation coefficient of 0.3. There are no correlations between the demands for products in different subgroups.

We compare two systems: (1) the sparse system (asymmetric equivalent of chaining) given in Figure 3.9 (which includes both primary and secondary links) whose cost of secondary production is $c_1$, and (2) the full flexibility system (i.e., any plant can produce any product) whose cost of secondary production is $c_2$. Assuming $p = 1$ and $c_p = 0$, our simulation study reveals that the sparse system with $c_1 = 0$ can beat full flexibility with $c_2 = 0.07$, and so on, as shown in Table 3.4. The numbers show that for this setting, upgrading to full flexibility provides very minimal benefits, that is, response is allowed to worsen by only a very small amount. Hence, our theory that system improvement must prioritize system response over system range seems to hold even under general asymmetric settings.

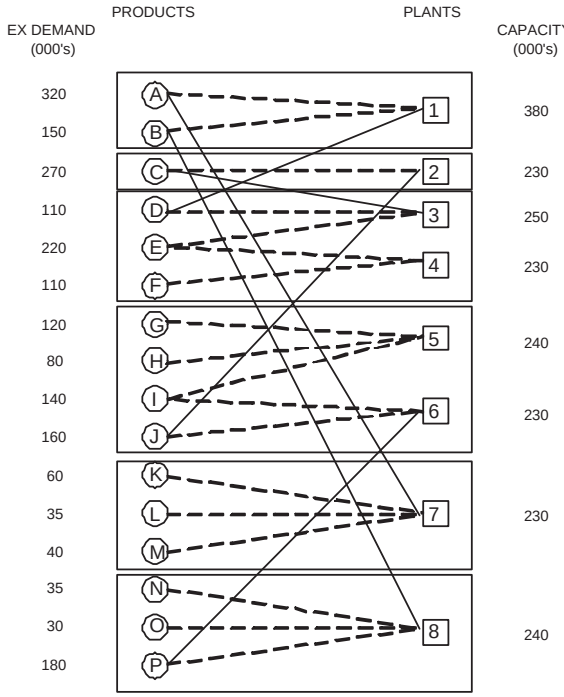


*Fig. 3.9:* Example of Asymmetric and Correlated System

| Secondary Cost | |
|---|---|
| Sparse System ($c_1$) | Full Flexibility ($c_2^*$) |
| 0.00 | 0.07 |
| 0.05 | 0.14 |
| 0.10 | 0.20 |
| 0.15 | 0.27 |
| 0.20 | 0.33 |

*Tab. 3.4:* Sparse System vs. Full Flexibility: Comparison of Secondary Production Costs (Asymmetric and Correlated System)

### 3.3.2 System Response and Demand Variability

The previous section clearly demonstrates the need to improve the response dimension of a system. This means reducing $c_s$ to the level of $c_p$. A natural question to ask is how much benefit does this bring? Using Theorem 5 and the random walk approach introduced in Chapter 2, we compare the asymptotic chaining efficiencies for high and low $c_s$ for some (discrete and continuous) uniform and normal distributions. It is easy to see that for a discrete uniform distribution with $2\Delta$ possible demand values, we have

$$\begin{aligned}
\mathrm{E}[\min(Y_1, Y_2)] &= \frac{(\Delta+1)(2\Delta+1)}{6\Delta^2} \cdot \mu \\
\mathrm{E}[Y_1] &= \frac{\Delta+1}{2\Delta} \cdot \mu \\
\mathrm{ACE}(c_s) &= \frac{\mathrm{E}[\min(Y_1, Y_2)]}{2\mathrm{E}[Y_1]} = \frac{2\Delta+1}{6\Delta} = \frac{1}{3} + \frac{1}{6\Delta} \quad \text{for high } c_s \\
\mathrm{ACE}(c_p) &= \frac{7\Delta+2}{12\Delta+6} = \frac{7}{12} - \frac{1}{8\Delta+4} \quad \text{when } c_s = c_p \\
\mathrm{CV} &= \sqrt{\frac{2\Delta^2+3\Delta+1}{6\Delta^2}}
\end{aligned}$$

We compile the results for some values of $\Delta$ in Table 3.5. Those for normal distributions are summarized in Table 3.6. These results suggest that production efficiency brings more benefits as the demand coefficient of variation decreases. This implies that although upgrading system response is important, it becomes even more so if coupled with initiatives to reduce demand uncertainty. In other words, we propose that improving production efficiency and reducing demand variability are complements.

| $\Delta$ | Distribution | CV | ACE($c_s$) for high $c_s$ | ACE($c_p$) | Improvement |
|---|---|---|---|---|---|
| 1 | 2-point | 1.00 | 0.5000 | 0.5000 | 0.0000 |
| 2 | 4-point | 0.79 | 0.4167 | 0.5333 | 0.1166 |
| 3 | 6-point | 0.72 | 0.3889 | 0.5476 | 0.1587 |
| 4 | 8-point | 0.68 | 0.3750 | 0.5556 | 0.1806 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\infty$ | continuous | 0.58 | 0.3333 | 0.5833 | 0.2500 |

*Tab. 3.5:* ACE Improvement for Upgrading System Response (Discrete Uniform Demand)

| CV | ACE($c_s$) for high $c_s$ | ACE($c_p$) | Improvement |
|---|---|---|---|
| 0.33 | 0.2929 | 0.7022 | 0.4093 |
| 0.31 | 0.2929 | 0.7145 | 0.4216 |
| 0.29 | 0.2929 | 0.7275 | 0.4346 |
| 0.27 | 0.2929 | 0.7413 | 0.4484 |
| 0.25 | 0.2929 | 0.7558 | 0.4629 |
| 0.23 | 0.2929 | 0.7708 | 0.4779 |
| 0.21 | 0.2929 | 0.7864 | 0.4935 |

*Tab. 3.6:* ACE Improvement for Upgrading System Response (Normal Demand)

# 4. VALUE OF THE THIRD CHAIN

In Section 2.4.4, we caught a glimpse of the potential value of employing a third chain in the design of flexible production systems. In the asymmetric case, we refer to the "third chain" as adding a third layer of flexibility (not necessarily a regular chain) to the sparse structure proposed by Chou et al. [16]. The optimal design of additional flexibility can be partially dealt with using either the constraint sampling methodology developed in [16] or the graph expander heuristics introduced in [17].

In this chapter, we further strengthen the case in support of the third chain in two ways: (1) when there is no full production postponement, and (2) when there are supply disruptions. We first consider the symmetric setting and demonstrate that the highly celebrated 2-chain is not sufficient in the presence of these two factors previously ignored in the literature. We further show that the 3-chain already recovers a bulk of the losses brought about by partial production postponement or supply disruptions, and that the 4-chain and higher chains can only provide minimal improvements.

For the asymmetric case, the interesting decision is how to design the "third chain". That is, given a budget to augment a third layer of flexibility, how does one choose from all the possible product-facility links? We apply the constraint sampling methodology introduced in [16] and discover

that, indeed, this additional amount of flexibility introduced into the system can already cushion most of the adverse effects of partial postponement and supply disruptions.

## 4.1 Process Flexibility and Production Postponement

With the advent of globalization and fragmentation of consumer demand, firms are faced with both uncertainty and complexity in their struggle to stay competitive in the global production and consumption network. Aside from process flexibility, another approach that can help deal with these challenges is "production postponement". Production postponement is "the firm's ability to set production quantities after demand uncertainty is resolved". When there is no postponement, the firm acts as a make-to-stock manufacturer; with full postponement, it behaves in a make-to-order fashion. In most instances, firms need to make simultaneous decisions on the level of flexibility as well as the level of postponement.

Clearly, the ideal solution is to have full flexibility and full postponement. However, full flexibility whereby all facilities can produce all products typically comes at great expense. Likewise, full postponement requires a highly responsive production system with short production lead times or the system may achieve poor service levels. It is precisely these concerns that prompt us as well as other papers in the literature to look into employing partial levels of flexibility and postponement. Assuming full postponement, Jordan and Graves [32] show that partial flexibility, in the form of a simple "chaining" strategy, can achieve nearly as much benefit as the full flexibility system

(almost 95%). Assuming full flexibility, Fisher and Raman [23] demonstrate that partial postponement using accurate response can lead to significant savings. Allowing a portion of capacity to be postponed to a point when some demand information is obtained, they found that for a major fashion skiwear company, cost relative to the existing system was reduced by enough to increase profits by 60%. However, to the best of our knowledge, none of the papers in the literature examine the benefits of implementing partial levels of **both** flexibility and postponement.

To better understand the importance of production postponement, we provide a brief review of the relevant literature. Inspired by the innovative practices of Benetton (Signorelli and Heskett [48]) and Hewlett-Packard (Lee et al [37]), the community has seen a growing interest in the study of production postponement. Alderson [4] appears to be the first to have coined the term "postponement" to refer to any strategy that allows for the "postponement of differentiation, such as postponing changes in form and identity to the latest possible point in the marketing flow or postponing change in inventory location to the latest possible point in time." Many subsequent works (e.g. Lee and Tang [38], Anand and Mendelson [5], and the references therein) studied the costs and benefits of postponement under different conditions. Interested readers may refer to the survey by Swaminathan and Lee [51].

In essence, one may see production postponement as a middle ground between make-to-stock and make-to-order systems. This is done by performing certain steps in the manufacturing process at an early period while postponing the remaining steps until demand uncertainty is resolved. An-

other way to connect the extremes of make-to-stock and make-to-order is by setting a portion of capacity to be consumed prior to knowledge of demand whereas the balance can be allocated to actual product demands as they become known. This is the model adopted in Fisher and Raman [23] as well as Van Mieghem and Dada [56]. The latter, however, focuses on the issue of comparing production postponement and price postponement, and does not consider the issue of flexibility which interests us in this paper. Nonetheless, it is the same way of modeling production postponement in those two papers that we also utilize in this paper. Our aim is to study the process flexibility problem under arbitrary levels of production postponement.

### *4.1.1 Model Description*

This section generalizes the process flexibility model under full postponement to the case where the postponement level can range anywhere between the extremes of make-to-stock and make-to-order. To this end, we develop a model to capture partial levels of both process flexibility and production postponement. The setting is as follows. We consider a system with $n$ plants and $n$ products. As before, we let $\mathcal{A}(n)$ and $\mathcal{B}(n)$ represent the set of product nodes and the set of plant nodes, respectively. The product demands are $\xi_1, \xi_2, \ldots, \xi_n$ which are independent and identically distributed random variables with distribution $F$ symmetrical around the mean $\mu$. This family of distributions includes the uniform and normal distributions. The plants, on the other hand, have fixed capacities of $\mu$ units each. We shall focus on this symmetric setting for the purpose of theory building and to gain basic

insights.

Early on, the firm carries out two strategic decisions; namely, the level of flexibility and the level of postponement. For flexibility, the firm chooses a flexibility configuration $\mathcal{G}(n) \subset \mathcal{A}(n) \times \mathcal{B}(n)$. We focus on symmetric flexibility structures and reduce the decision to a scalar $d$, denoting the common node degree. Moreover, we consider structures that form the longest chains possible due to their well-established efficiency (see [32], [16]). That is, we consider $d$-chains as defined in Section 1.3.

For production postponement, we model a two-period production process and define $\alpha$ as the proportion of capacity postponed to the second period while $1 - \alpha$ is for first-period consumption. When $\alpha = 0$, we have a make-to-stock setting and all production must be decided in the first period. When $\alpha = 1$, our model reduces to the make-to-order, full-postponement setting in the literature. We allow the firm to choose its desired postponement level $\alpha$ over the range $[0, 1]$.

Given any combination of $\mathcal{G}(n)$ (equivalently, $d$) and $\alpha$, the expected mismatch cost can be determined by solving the following two-stage problem. In the first stage, $(1-\alpha)\mu$ units are made available at each plant to produce whatever allowed combination of products $1, 2, \ldots, n$ to stock, i.e. without information on actual final demand. In the second stage, the remaining $\alpha\mu$ units in each plant become available to meet whatever actual demand the firm cannot fill from first-stage stock. Our problem here is essentially a multi-item newsvendor model with second-stage supply and partial capacity sharing, which we refer to as the Minimum Mismatch Cost Model in Section 1.3.1. In our analysis, we assume that overstocking and understocking are

equally penalized, i.e. $c_o = c_u$.

$$
\begin{aligned}
(P1): \quad G^*_{\mathcal{G}(n)}(\alpha) = \; & \min_{\boldsymbol{x}} \; G_{\mathcal{G}(n)}(\boldsymbol{x}, \alpha) \\
& \text{s.t.} \quad \sum_{i=1}^{n} x_{ij} \leq (1-\alpha)\mu \;\; \forall j = 1, 2, \ldots n \\
& \qquad x_{ij} \geq 0 \;\; \forall i, j = 1, 2, \ldots n \\
& \qquad x_{ij} = 0 \;\; \forall (i,j) \notin \mathcal{G}(n)
\end{aligned}
$$

where

$$G_{\mathcal{G}(n)}(\boldsymbol{x}, \alpha) = g_1(\boldsymbol{x}) + g_2(\boldsymbol{x}) - \mathrm{E}[h_{\mathcal{G}(n)}(\boldsymbol{x}, \alpha, \boldsymbol{\xi})]$$

$$g_1(\boldsymbol{x}) = \sum_{i=1}^{n} \int_0^{\sum_{j=1}^{n} x_{ij}} \left( \sum_{j=1}^{n} x_{ij} - \xi_i \right) dF(\xi_i)$$

$$g_2(\boldsymbol{x}) = \sum_{i=1}^{n} \int_{\sum_{j=1}^{n} x_{ij}}^{\infty} \left( \xi_i - \sum_{j=1}^{n} x_{ij} \right) dF(\xi_i)$$

and

$$
\begin{aligned}
h_{\mathcal{G}(n)}(\boldsymbol{x}, \alpha, \boldsymbol{\xi}) = \; & \max_{\boldsymbol{y}} \; \sum_{i=1}^{n} \sum_{j=1}^{n} y_{ij} \\
& \text{s.t.} \quad \sum_{j=1}^{n} y_{ij} \leq \left( \xi_i - \sum_{j=1}^{n} x_{ij} \right)^+ \;\; \forall i = 1, 2, \ldots n \\
& \qquad \sum_{i=1}^{n} y_{ij} \leq \alpha\mu \;\; \forall j = 1, 2, \ldots n \\
& \qquad y_{ij} \geq 0 \;\; \forall i, j = 1, 2, \ldots n \\
& \qquad y_{ij} = 0 \;\; \forall (i,j) \notin \mathcal{G}(n)
\end{aligned}
$$

Allowing first-stage production to be fully flexible while holding second-

stage production to $\mathcal{G}(n)$-flexibility, we have Problem 2 as follows. Notice that under full flexibility, there will be multiple optimal solutions. Hence, the $n^2$-dimensional decision vector $\boldsymbol{x}$ can be reduced to the $n$-dimensional decision vector $\boldsymbol{z}$ by letting $z_i = \sum_{j=1}^{n} x_{ij}, \forall i = 1, 2, \ldots, n$.

$$
\begin{aligned}
(P2): \quad \overline{G}^*_{\mathcal{G}(n)}(\alpha) = \quad & \min_{\boldsymbol{x}} \quad G_{\mathcal{G}(n)}(\boldsymbol{x}, \alpha) \\
& \text{s.t.} \quad \sum_{i=1}^{n} x_{ij} \leq (1-\alpha)\mu \quad \forall j = 1, 2, \ldots n \\
& \qquad\quad x_{ij} \geq 0 \quad \forall i, j = 1, 2, \ldots n \\
= \quad & \min_{\boldsymbol{z}} \quad \overline{G}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha) \\
& \text{s.t.} \quad \sum_{i=1}^{n} z_i \leq (1-\alpha)n\mu \\
& \qquad\quad z_i \geq 0 \quad \forall i = 1, 2, \ldots n
\end{aligned}
$$

where

$$\overline{G}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha) = \overline{g}_1(\boldsymbol{z}) + \overline{g}_2(\boldsymbol{z}) - \mathrm{E}[\overline{h}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha, \boldsymbol{\xi})]$$

$$\overline{g}_1(\boldsymbol{z}) = \sum_{i=1}^{n} \int_0^{z_i} (z_i - \xi_i) dF(\xi_i)$$

$$\overline{g}_2(\boldsymbol{z}) = \sum_{i=1}^{n} \int_{z_i}^{\infty} (\xi_i - z_i) dF(\xi_i)$$

and

$$\begin{aligned}
\overline{h}_{\mathcal{G}(n)}(\boldsymbol{z},\alpha,\boldsymbol{\xi}) = \quad \max_{\boldsymbol{y}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij} \\
\text{s.t.} \quad & \sum_{j=1}^{n} y_{ij} \le (\xi_i - z_i)^+ \quad \forall i = 1,2,\dots n \\
& \sum_{i=1}^{n} y_{ij} \le \alpha\mu \quad \forall j = 1,2,\dots n \\
& y_{ij} \ge 0 \quad \forall i,j = 1,2,\dots n \\
& y_{ij} = 0 \quad \forall (i,j) \notin \mathcal{G}(n)
\end{aligned}$$

### 4.1.2 Insufficiency of the 2-Chain

In this section, we compare the optimal expected mismatch cost of the chaining strategy with that of the full flexibility system. In order to achieve that, we must first characterize the optimal solution $\boldsymbol{x}^*$ of (P1) and (P2). The theory of majorization and Schur-convexity (see [42]) is utilized as follows. We let $x_{(1)}$ indicate the largest element in vector $\boldsymbol{x}$, $x_{(2)}$ indicate the second-largest element, and so on.

**Definition 1.** *The vector $\boldsymbol{x}$ is said to majorize the vector $\boldsymbol{y}$ (denoted $\boldsymbol{x} \succ \boldsymbol{y}$) if*

$$\sum_{i=1}^{k} x_{(i)} \ge \sum_{i=1}^{k} y_{(i)} \qquad \forall k = 1,2,\dots,n-1$$

$$\text{and} \quad \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

**Definition 2.** *A function $f : \mathcal{R}^n \to \mathcal{R}$ is called Schur-convex if*

$$\boldsymbol{x} \succ \boldsymbol{y} \Rightarrow f(\boldsymbol{x}) \geq f(\boldsymbol{y})$$

**Lemma 6.** *Suppose $f : \mathcal{R}^n \to \mathcal{R}$ and dom $f = \mathcal{R}^n_+$. Define $g : \mathcal{R}^n \to \mathcal{R}$ by $g(\boldsymbol{x}) = f(\boldsymbol{x}^+)$, where $\boldsymbol{x}^+$ is the component-wise positive part of $\boldsymbol{x}$. If $f$ is convex in $\boldsymbol{x}$ and nondecreasing in each argument $x_i$ over $[0, \infty)$, then $g$ is convex in $\boldsymbol{x}$.*

**Proof.** Without loss of generality, let $\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \boldsymbol{x}_3 \\ \boldsymbol{x}_4 \end{pmatrix}, \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \boldsymbol{y}_3 \\ \boldsymbol{y}_4 \end{pmatrix} \in \mathcal{R}^n$ such that $\boldsymbol{x}_1 \geq 0, \boldsymbol{y}_1 \geq 0, \boldsymbol{x}_2 \geq 0, \boldsymbol{y}_2 \leq 0, \boldsymbol{x}_3 \leq 0, \boldsymbol{y}_3 \geq 0, \boldsymbol{x}_4 \leq 0, \boldsymbol{y}_4 \leq 0; \boldsymbol{x}_i, \boldsymbol{y}_i \in \mathcal{R}^{n_i}, \forall i = 1, 2, 3, 4$ and $\sum_{i=1}^{4} n_i = n$. For $\lambda \in [0, 1]$,

$$
\begin{aligned}
g\begin{pmatrix} \lambda\boldsymbol{x}_1+(1-\lambda)\boldsymbol{y}_1 \\ \lambda\boldsymbol{x}_2+(1-\lambda)\boldsymbol{y}_2 \\ \lambda\boldsymbol{x}_3+(1-\lambda)\boldsymbol{y}_3 \\ \lambda\boldsymbol{x}_4+(1-\lambda)\boldsymbol{y}_4 \end{pmatrix} &= f\begin{pmatrix} \lambda\boldsymbol{x}_1+(1-\lambda)\boldsymbol{y}_1 \\ (\lambda\boldsymbol{x}_2+(1-\lambda)\boldsymbol{y}_2)^+ \\ (\lambda\boldsymbol{x}_3+(1-\lambda)\boldsymbol{y}_3)^+ \\ 0 \end{pmatrix} \\
&\leq f\begin{pmatrix} \lambda\boldsymbol{x}_1+(1-\lambda)\boldsymbol{y}_1 \\ \lambda\boldsymbol{x}_2 \\ (1-\lambda)\boldsymbol{y}_3 \\ 0 \end{pmatrix} \\
&\leq \lambda f\begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ 0 \\ 0 \end{pmatrix} + (1-\lambda) f\begin{pmatrix} \boldsymbol{y}_1 \\ 0 \\ \boldsymbol{y}_3 \\ 0 \end{pmatrix} \\
&= \lambda g\begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \boldsymbol{x}_3 \\ \boldsymbol{x}_4 \end{pmatrix} + (1-\lambda) g\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \boldsymbol{y}_3 \\ \boldsymbol{y}_4 \end{pmatrix}
\end{aligned}
$$

The first inequality holds because $\lambda\boldsymbol{x}_2 \geq \lambda\boldsymbol{x}_2+(1-\lambda)\boldsymbol{y}_2; \lambda\boldsymbol{x}_2 \geq 0; (1-\lambda)\boldsymbol{y}_3 \geq \lambda\boldsymbol{x}_3+(1-\lambda)\boldsymbol{y}_3; (1-\lambda)\boldsymbol{y}_3 \geq 0$ and $f$ is increasing in each argument. The second inequality is due to the convexity of $f$ while the equations follow from definition. ■

**Proposition 1.** *$\overline{G}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha)$ is Schur-convex for any symmetric structure $\mathcal{G}(n)$.*

**Proof.** According to Marshall and Olkin [42], it suffices to show that $\overline{G}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha)$ is symmetric and convex in $\boldsymbol{z}$. Clearly, $\overline{g}_1(\boldsymbol{z})$ and $\overline{g}_2(\boldsymbol{z})$ are symmetric, i.e. any two of its arguments can be swapped without modifying the function value. For symmetric structure $\mathcal{G}(n)$, $\mathrm{E}[\overline{h}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha, \boldsymbol{\xi})]$ is also symmetric. Hence, $\overline{G}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha)$ is symmetric.

Next, we can see that $\overline{g}_1(\boldsymbol{z})$ is separable and each individual term $\overline{g}_{1i}(z_i) = \int_0^{z_i}(z_i - \xi_i)dF(\xi_i)$ is convex because $\overline{g}''_{1i}(z_i) = f(z_i) \geq 0$. Therefore, $\overline{g}_1(\boldsymbol{z})$ is convex. We then turn our attention to

$$\overline{g}_2(\boldsymbol{z}) - \mathrm{E}[\overline{h}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha, \boldsymbol{\xi})] = \mathrm{E}\left[\sum_{i=1}^{n}(\xi_i - z_i)^+ - \overline{h}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha, \boldsymbol{\xi})\right]$$

Letting $b_i = (\xi_i - z_i)^+$, we want to characterize the convexity of $\sum_{i=1}^{n} b_i - \overline{h}_{\mathcal{G}(n)}(\boldsymbol{b}, \alpha)$ where

$$\begin{aligned}
\overline{h}_{\mathcal{G}(n)}(\boldsymbol{b}, \alpha) = \quad \max_{\boldsymbol{y}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij} \\
\text{s.t.} \quad & \sum_{j=1}^{n} y_{ij} \leq b_i \quad \forall i = 1, 2, \ldots n \\
& \sum_{i=1}^{n} y_{ij} \leq \alpha\mu \quad \forall j = 1, 2, \ldots n \\
& y_{ij} \geq 0 \quad \forall i, j = 1, 2, \ldots n \\
& y_{ij} = 0 \quad \forall (i, j) \notin \mathcal{G}(n)
\end{aligned}$$

According to Bertsimas and Tsitsiklis [8], $\overline{h}_{\mathcal{G}(n)}(\boldsymbol{b}, \alpha)$ is concave in $\boldsymbol{b}$. It is

also easy to see that $\sum_{i=1}^{n} b_i - \overline{h}_{\mathcal{G}(n)}(\boldsymbol{b}, \alpha) \geq 0$ and $0 \leq \frac{\partial \overline{h}_{\mathcal{G}(n)}}{\partial b_i} \leq 1$. It follows that $\sum_{i=1}^{n} b_i - \overline{h}_{\mathcal{G}(n)}(\boldsymbol{b}, \alpha)$ is convex in $\boldsymbol{b}$ and nondecreasing in each argument $b_i$ over $[0, \infty)$. By Lemma 6 and preservation of convexity under composition with affine function and expectation, $\overline{g}_2(\boldsymbol{z}) - \mathrm{E}[\overline{h}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha, \boldsymbol{\xi})]$ is convex in $\boldsymbol{z}$. Hence, $\overline{G}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha)$ is convex in $\boldsymbol{z}$. ■

**Lemma 7.** *Suppose $f : \mathcal{R} \to \mathcal{R}$ is an increasing convex function while $g, \hat{g} : \mathcal{R} \to \mathcal{R}$ are decreasing convex functions such that $\hat{g}'(x) < g'(x) \leq 0$. If $x^*$ minimizes $f(x) + g(x)$ and $\hat{x}^*$ minimizes $f(x) + \hat{g}(x)$, then $x^* \leq \hat{x}^*$.*

**Proof.** It follows from optimality that $f'(x^*) = -g'(x^*)$ and $f'(\hat{x}^*) = -\hat{g}'(\hat{x}^*)$. Since $f$ is convex while $-g, -\hat{g}$ are concave, $f'$ is nondecreasing and $-g', -\hat{g}'$ are nonincreasing. Because $-\hat{g}'(x) > -g'(x)$, $x^* \leq \hat{x}^*$. ■

**Proposition 2.** *$x^*_{ii} = (1 - \alpha)\mu, \forall i = 1, 2, \ldots, n$ and $x^*_{ij} = 0, \forall i \neq j$ is a solution to both (P1) and (P2).*

**Proof.** Consider first (P2). From Proposition 1 and Marshall and Olkin [42], the Schur-convex function $\overline{G}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha)$ is minimized at $z_i^* = z_0, \forall i = 1, 2, \ldots, n$. We want to find $z_0$ that minimizes

$$\begin{aligned} \overline{G}_{\mathcal{G}(n)}(z_0\mathbf{1}, \alpha) &= \overline{g}_1(z_0\mathbf{1}) + \overline{h}_1(z_0\mathbf{1}, \alpha) \\ &\geq \overline{g}_1(z_0\mathbf{1}) + \overline{h}_2(z_0\mathbf{1}, \alpha) \\ &\geq \overline{g}_1(z_0\mathbf{1}) + \overline{h}_3(z_0\mathbf{1}, \alpha) \end{aligned}$$

where

$$\begin{aligned}
\overline{h}_1(z_0\mathbf{1},\alpha) &= \mathrm{E}\left[\sum_{i=1}^{n}(\xi_i - z_0)^+ - \overline{h}_{\mathcal{G}(n)}(z_0\mathbf{1},\alpha,\boldsymbol{\xi})\right] \\
\overline{h}_2(z_0\mathbf{1},\alpha) &= \mathrm{E}\left[\sum_{i=1}^{n}(\xi_i - z_0)^+ - \overline{h}_{\mathcal{F}(n)}(z_0\mathbf{1},\alpha,\boldsymbol{\xi})\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{n}(\xi_i - z_0)^+ - \min\left(\sum_{i=1}^{n}(\xi_i - z_0)^+, \alpha n\mu\right)\right] \\
&= \mathrm{E}\left[\max\left(0, \sum_{i=1}^{n}(\xi_i - z_0)^+ - \alpha n\mu\right)\right] \\
\overline{h}_3(z_0\mathbf{1},\alpha) &= \mathrm{E}\left[\max\left(0, \sum_{i=1}^{n}(\xi_i - z_0) - \alpha n\mu\right)\right]
\end{aligned}$$

Let

$$\begin{aligned}
\hat{G}(z_0) &= \overline{g}_1(z_0\mathbf{1}) + \overline{h}_3(z_0\mathbf{1},\alpha) \\
&= \sum_{i=1}^{n}\int_0^{z_0}(z_0 - \xi_i)dF(\xi_i) + \int_{nz_0+\alpha n\mu}^{\infty}(\xi - nz_0 - \alpha n\mu)d\hat{F}(\xi)
\end{aligned}$$

where $\xi = \sum_{i=1}^{n}\xi_i \sim \hat{F}$.

To minimize $\hat{G}(z_0)$ such that $z_0 \leq (1-\alpha)\mu$, we take the derivative as follows.

$$\begin{aligned}
\hat{G}'(z_0) &= nF(z_0) - n[1 - \hat{F}(nz_0 + \alpha n\mu)] \\
&= n[F(z_0) + \hat{F}(nz_0 + \alpha n\mu) - 1] \\
&\leq n\left[\frac{1}{2} + \hat{F}(n\mu) - 1\right], \quad \forall z_0 \leq (1-\alpha)\mu \\
&= n\left[\frac{1}{2} + \frac{1}{2} - 1\right] = 0, \quad \forall z_0 \leq (1-\alpha)\mu
\end{aligned}$$

That $\hat{G}'((1-\alpha)\mu) \leq 0$ implies that the unconstrained solution $\hat{z}_0^* \geq (1-\alpha)\mu$. Moreover, it can be shown that $\frac{\partial \overline{h}_1}{\partial z_0} < \frac{\partial \overline{h}_2}{\partial z_0} < \frac{\partial \overline{h}_3}{\partial z_0} \leq 0$. By Lemma 7, the unconstrained minimizer of $\overline{G}_{\mathcal{G}(n)}(z_0\mathbf{1}, \alpha)$ is $z_0^* \geq \hat{z}_0^* \geq (1-\alpha)\mu$. Hence, the optimal solution to (P2) is $z_i^* = (1-\alpha)\mu, \forall i = 1, 2, \ldots, n$. This is equivalent to $x_{ii}^* = (1-\alpha)\mu, \forall i = 1, 2, \ldots, n$ and $x_{ij}^* = 0, \forall i \neq j$. Since this solution is also feasible for (P1) and the feasible set of (P1) is a subset of the feasible set of (P2), it also solves (P1). ■

Proposition 2 tells us that under the given conditions, the optimal first-stage production is to exhaust all first-stage capacity for primary production regardless of the flexibility structure. This implies that when $\alpha = 0$, i.e. there is no postponement of production, any form of flexibility brings no additional benefits. In fact, when no postponement is possible, it may be worthwhile to consider a dedicated system ($d = 1$) with no flexibility at all. Of course, if the firm anticipates an improvement in postponement in the future, existing flexibility should not be uninstalled.

That said, we want to characterize the performance gap between full flexibility and the 2-chain as the level of postponement increases. As in Section 1.3, we let $\mathcal{C}(n)$ denote the 2-chain structure and define

$$\begin{aligned} \Delta G(\alpha) &\stackrel{\Delta}{=} G^*_{\mathcal{C}(n)}(\alpha) - G^*_{\mathcal{F}(n)}(\alpha) \\ &= \mathrm{E}[\overline{h}_{\mathcal{F}(n)}((1-\alpha)\mu\mathbf{1}, \alpha, \boldsymbol{\xi})] - \mathrm{E}[\overline{h}_{\mathcal{C}(n)}((1-\alpha)\mu\mathbf{1}, \alpha, \boldsymbol{\xi})] \end{aligned}$$

where the second equation is due to Proposition 2, which allows $g_1(\boldsymbol{x}^*) + g_2(\boldsymbol{x}^*)$ to cancel out.

**Proposition 3.** *$\exists\alpha \in (0,1)$ such that $\Delta G(\alpha)$ is largest.*

**Proof.** Define $\hat{h}_{\mathcal{G}(n)}(\alpha) = \mathrm{E}[\overline{h}_{\mathcal{G}(n)}((1-\alpha)\mu\mathbf{1}, \alpha, \boldsymbol{\xi})]$ and

$$\begin{aligned}
\overline{h}_{\mathcal{G}(n)}((1-\alpha)\mu\mathbf{1}, \alpha, \boldsymbol{\xi}) = \quad & \max_{\boldsymbol{y}} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij} \\
& \text{s.t.} \quad \sum_{j=1}^{n} y_{ij} \le (\xi_i - (1-\alpha)\mu)^+ \quad \forall i = 1, 2, \ldots n \\
& \qquad \sum_{i=1}^{n} y_{ij} \le \alpha\mu \quad \forall j = 1, 2, \ldots n \\
& \qquad y_{ij} \ge 0 \quad \forall i, j = 1, 2, \ldots n \\
& \qquad y_{ij} = 0 \quad \forall (i,j) \notin \mathcal{G}(n)
\end{aligned}$$

so that $\Delta G(\alpha) = \hat{h}_{\mathcal{F}(n)}(\alpha) - \hat{h}_{\mathcal{C}(n)}(\alpha)$.

It can also be shown that $\hat{h}_{\mathcal{F}(n)}(\alpha) \ge \hat{h}_{\mathcal{C}(n)}(\alpha), \hat{h}_{\mathcal{F}(n)}(0) = \hat{h}_{\mathcal{C}(n)}(0)$, and $\lim_{\alpha\to\infty} \hat{h}_{\mathcal{F}(n)}(\alpha) = \lim_{\alpha\to\infty} \hat{h}_{\mathcal{C}(n)}(\alpha)$. According to Bertsimas and Tsitsiklis [8], $\hat{h}_{\mathcal{G}(n)}(\alpha)$ is increasing and concave in $\alpha$ for any $\mathcal{G}(n)$. It follows that $\exists\alpha \in (0,\infty)$ such that $\Delta G(\alpha)$ is largest. For $\alpha \ge 1$,

$$\begin{aligned}
\hat{h}_{\mathcal{F}(n)}(\alpha) &= \mathrm{E}\left[\min\left(\sum_{i=1}^{n}(\xi_i - (1-\alpha)\mu)^+, \alpha n\mu\right)\right] \\
&= \mathrm{E}\left[\min\left(\sum_{i=1}^{n}(\xi_i - \mu + \alpha\mu), \alpha n\mu\right)\right] \\
&= \alpha n\mu + \mathrm{E}\left[\min\left(\sum_{i=1}^{n}(\xi_i - \mu), 0\right)\right]
\end{aligned}$$

Since $\hat{h}_{\mathcal{F}(n)}(\alpha)$ is increasing and linear in $\alpha$ over $[1,\infty)$, $\hat{h}_{\mathcal{C}(n)}(\alpha)$ is increasing and concave in $\alpha$, $\hat{h}_{\mathcal{F}(n)}(1) > \hat{h}_{\mathcal{C}(n)}(1)$ (see Chou et al [16]), and $\lim_{\alpha\to\infty} \hat{h}_{\mathcal{F}(n)}(\alpha) =$

$\lim_{\alpha\to\infty} \hat{h}_{\mathcal{C}(n)}(\alpha)$, it follows that $\Delta G(\alpha)$ is decreasing in $\alpha$ over $[1, \infty)$. Hence, $\exists \alpha \in (0, 1)$ such that $\Delta G(\alpha)$ is largest. ■

Proposition 3 suggests that for certain levels of partial postponement, the performance gap between full flexibility and the 2-chain may not be as small as it is under the full postponement case. In a make-to-order environment ($\alpha = 1$ ), it is already known in the literature that the 2-chain performs almost as well as full flexibility. On the other hand, in a make-to-stock scenario ($\alpha = 0$ ), any form of flexibility is of no benefit, thus the 2-chain and full flexibility would incur the same cost. For some $\alpha \in (0, 1)$ though, the difference between full flexibility and 2-chain may be quite significant.

That said, if one wants to approximate the benefits of full flexibility and full postponement using only partial levels of both these dimensions, care has to be taken in choosing the proper levels of flexibility and postponement that can give the desired result. Unlike the process flexibility literature which finds that minimal partial flexibility is enough to achieve the benefits of the first-best solution, it may not be true with just any low levels of flexibility and postponement. In the event of partial postponement, more flexibility is necessary to make up for the performance loss. The question becomes how much additional flexibility is enough.

### *4.1.3 Sufficiency of the 3-Chain*

In reality, every firm has to contend with the limitation that it can only employ partial levels of both production postponement and process flexibility.

However, this does not stop firms from trying to find postponement-flexibility configurations that achieve a high percentage of the performance of full flexibility and full postponement (which we call the first-best solution). With full postponement, a 2-chain is enough to approximate the performance of the first-best solution. Without full postponement, we have shown that the 2-chain does not suffice. The performance loss can be attributed to two factors: flexibility loss and postponement loss. Under full postponement, there is no postponement loss and the 2-chain incurs minimal flexibility loss. Under partial postponement, postponement loss is incurred but on top of that, the flexibility loss of the 2-chain also becomes quite significant. This explains why, for example, a 2-chain with 50% postponement can perform quite badly.

While using additional flexibility to cushion the postponement loss is interesting and important, we shall defer this discussion to the next section. Here, we focus first on how to reduce the flexibility loss when there is partial postponement. To do so, we use the method of asymptotic analysis introduced in Chapter 2. For every pair of degree of flexibility $d$ and level of postponement $\alpha$, the flexibility efficiency (the inverse of flexibility loss) can be measured as in Section 1.3.2.

$$\begin{aligned} ACE_d(\alpha) &= \lim_{n\to\infty} \frac{G^*_{\mathcal{D}(n)}(\alpha) - G^*_{\mathcal{C}_d(n)}(\alpha)}{G^*_{\mathcal{D}(n)}(\alpha) - G^*_{\mathcal{F}(n)}(\alpha)} \\ &= \lim_{n\to\infty} \frac{\hat{h}_{\mathcal{C}_d(n)}(\alpha) - \hat{h}_{\mathcal{D}(n)}(\alpha)}{\hat{h}_{\mathcal{F}(n)}(\alpha) - \hat{h}_{\mathcal{D}(n)}(\alpha)} \end{aligned}$$

where $\hat{h}_{\mathcal{G}(n)}(\alpha) = \mathrm{E}[\overline{h}_{\mathcal{G}(n)}((1-\alpha)\mu\mathbf{1}, \alpha, \boldsymbol{\xi})]$ and

$$\begin{aligned}
\overline{h}_{\mathcal{G}(n)}((1-\alpha)\mu\mathbf{1}, \alpha, \boldsymbol{\xi}) = \quad & \max_{\boldsymbol{y}} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij} \\
& \text{s.t.} \quad \sum_{j=1}^{n} y_{ij} \leq (\xi_i - (1-\alpha)\mu)^+ \quad \forall i = 1, 2, \ldots n \\
& \qquad \sum_{i=1}^{n} y_{ij} \leq \alpha\mu \quad \forall j = 1, 2, \ldots n \\
& \qquad y_{ij} \geq 0 \quad \forall i, j = 1, 2, \ldots n \\
& \qquad y_{ij} = 0 \quad \forall (i,j) \notin \mathcal{G}(n)
\end{aligned}$$

Observe that $\overline{h}_{\mathcal{G}(n)}$ is an instance of the Maximum Flow Model described in Section 1.3.1, with capacities $C_j = \alpha\mu$ and identical distributed demand random variables $D_i = (\xi_i - (1-\alpha)\mu)^+$. Note that this demand distribution is just a truncated version of the original distribution with negative drift. The resulting system will no longer be symmetric because capacity does not equal expected demand and demand is not symmetrical around its mean. Nonetheless, the method using alternating renewal process initiated in Section 2.4.1, as well as its extension to higher-degree chains in Section 2.4.4, can be used to study our problem at hand.

Consider demand that follows a normal distribution with a coefficient of variation of 0.30, and total capacity that equals expected demand. Table 4.1 and Figure 4.1 summarize the asymptotic chaining efficiency for various levels of production postponement and partial flexibility. Under full postponement, we already expect the 2-chain to perform quite well providing 72% of the benefits of full flexibility even if the system size becomes ridicu-

| $\alpha$ | Flexibility Structure | | | |
|---|---|---|---|---|
| | 2-chain | 3-chain | 4-chain | 5-chain |
| 0.1 | 18.29% | 26.73% | 31.04% | 33.38% |
| 0.2 | 31.99% | 45.24% | 51.89% | 55.67% |
| 0.3 | 42.19% | 57.62% | 65.11% | 69.44% |
| 0.4 | 49.88% | 65.88% | 73.38% | 77.69% |
| 0.5 | 55.80% | 71.52% | 78.64% | 82.68% |
| 0.6 | 60.48% | 75.52% | 82.12% | 85.82% |
| 0.7 | 64.26% | 78.48% | 84.55% | 87.91% |
| 0.8 | 67.38% | 80.76% | 86.34% | 89.39% |
| 0.9 | 70.00% | 82.58% | 87.72% | 90.51% |
| 1.0 | 72.23% | 84.08% | 88.84% | 91.40% |

*Tab. 4.1:* Asymptotic Chaining Efficiency for Various Levels of Production Postponement and Partial Flexibility

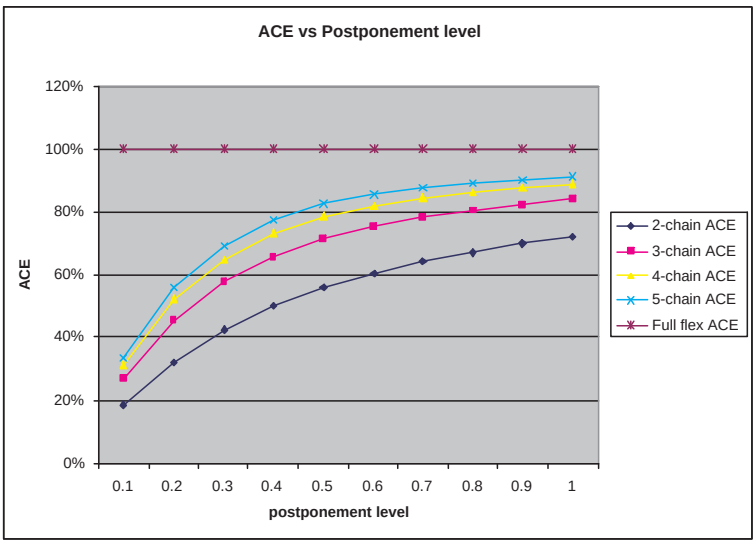


*Fig. 4.1:* Asymptotic Chaining Efficiency vs Level of Production Postponement

lously large, say 1 million × 1 million. However, under 50% postponement, this number drops to only 55%. This confirms our earlier result that the 2-chain may not be sufficient under partial postponement. Fortunately, adding a third chain can restore the performance back to almost 72%. Adding a fourth chain can bring some benefits but significantly less than the gain from 2-chain to 3-chain. We also see that further improvements from the fifth and higher chains are negligible. Such investments are no longer worthwhile, and especially so in the rather common scenario where the cost of additional flexibility increases in the amount of flexibility already installed. Notice also that for low postponement levels (say 10%), no amount of partial flexibility may be able to recover the flexibility loss in the system, unless one considers the enormous investment in full flexibility. For various other scenarios (different coefficient of variation and different demand distributions), we also report similar results – that in the case of partial postponement, the 2-chain is insufficient, but the 3-chain is enough. This observation contrasts with the process flexibility literature which believes that the additional benefit from using the third chain is negligible.

### *4.1.4 The Flexibility-Postponement Trade-off*

In this section, we turn our emphasis to overall performance loss, which includes both flexibility loss and postponement loss. To this end, we examine the total expected mismatch cost as it changes with respect to various levels of postponement and flexibility. We also explore the interesting trade-off between process flexibility and production postponement. That is, for a

given target performance level (e.g. mismatch cost), what are the different combinations of flexibility and postponement which can achieve this target? Such analysis can aid in the proper allocation of resources between flexibility and postponement.

We conduct a numerical study on a 10-plant, 10-product system. Each product has demand that follows a normal distribution with mean 2000 units and standard deviation 600 units. Each plant has a capacity of 2000 units. For each postponement level $\alpha \in \{0.00, 0.05, 0.10, \ldots, 0.95, 1.00\}$ and each degree of flexibility $d \in \{1, 2, 3, \ldots, 9, 10\}$, we computed the minimum mismatch cost over a set of 1000 demand scenarios by solving the following optimization problem. This problem, which is equivalent to problem (P1), can then be formulated as a large linear program.

$$
\begin{aligned}
\min_{\boldsymbol{x},\boldsymbol{y}} \quad & \frac{1}{1000}\sum_{k=1}^{1000}\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{n} x_{ij} + \sum_{j=1}^{n} y_{ij}^{k} - \xi_i^k\right)^{+} + \left(\xi_i^k - \sum_{j=1}^{n} x_{ij} - \sum_{j=1}^{n} y_{ij}^{k}\right)^{+}\right] \\
\text{s.t.} \quad & \sum_{i=1}^{n} x_{ij} \leq (1-\alpha)\mu \quad \forall j = 1, 2, \ldots n \\
& \sum_{i=1}^{n} y_{ij}^{k} \leq \alpha\mu \quad \forall j = 1, 2, \ldots n, \forall k = 1, 2, \ldots, 1000 \\
& x_{ij} \geq 0 \quad \forall i, j = 1, 2, \ldots n \\
& y_{ij}^{k} \geq 0 \quad \forall i, j = 1, 2, \ldots n, \forall k = 1, 2, \ldots, 1000 \\
& x_{ij} = 0 \quad \forall (i,j) \notin \mathcal{G}(n) \\
& y_{ij}^{k} = 0 \quad \forall (i,j) \notin \mathcal{G}(n), \forall k = 1, 2, \ldots, 1000
\end{aligned}
$$

where $\xi_i^k$ is the demand for product $i$ under the $k$th demand scenario, while $y_{ij}^k$ is the second-stage production allocation upon seeing the demand scenario.

We plot the expected mismatch cost against the postponement level for different levels of flexibility, as well as the expected mismatch cost against the level of flexibility for different postponement levels. The respective graphs are shown in Figure 4.2 and Figure 4.3. Observe that the expected cost is decreasing and convex in the level of either dimension (flexibility or postponement), implying their diminishing values. This tells us that a little bit of either dimension can already bring about substantial benefits. However, we also see that the rate at which the cost diminishes is increasing in the other dimension – postponement and flexibility are complements. When one dimension is low, limited improvement in the other dimension may not bring as much benefits and so more increase in this second dimension will continue to generate value.

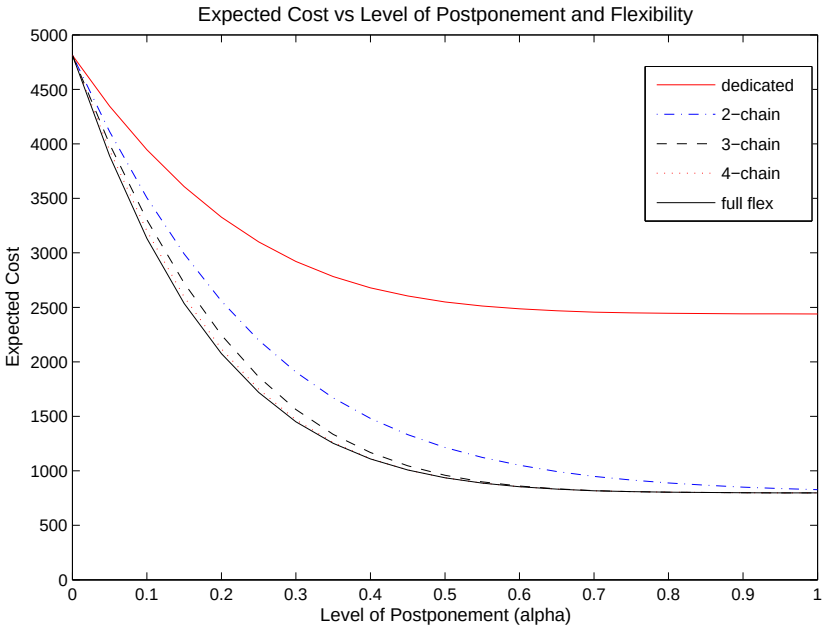


*Fig. 4.2:* Expected Mismatch Cost vs. Level of Production Postponement

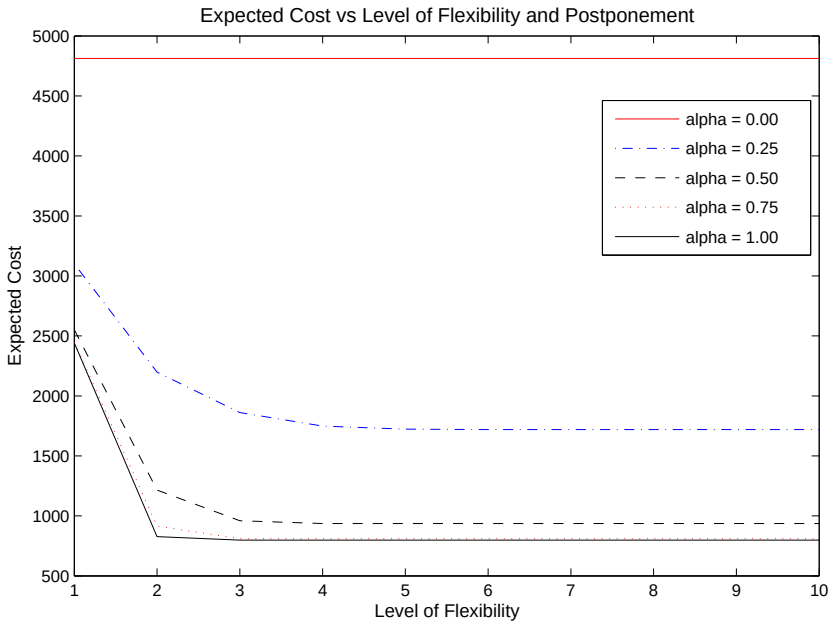


*Fig. 4.3:* Expected Mismatch Cost vs. Level of Process Flexibility

From Figure 4.2, we also confirm the following earlier findings.

1. With no postponement, any form of flexibility is useless. At $\alpha = 0$, all five lines converge to the same point.

2. With full postponement, 2-chains and up are very efficient. At $\alpha = 1$, see the huge gap between the dedicated system and full flexibility, whereas the 2-chain line is very near full flexibility.

3. Under partial postponement, 2-chain may not be sufficient. For $\alpha \in [0.1, 0.7]$, the gap between 2-chain and full flexibility is quite sizable, especially between 0.2 and 0.5.

4. Under partial postponement, 3-chain and up continues to be quite efficient. One can clearly see that the 3-chain line very nearly traces the

full flexibility line, recovering most of the flexibility loss at all postponement levels.

Furthermore, the computational data we obtained can be used to demonstrate the trade-off between flexibility and postponement. For a given level of mismatch cost, we plot into a curve the different combinations of flexibility and postponement levels that return that cost level. We do the same for several different expected cost levels and come up with the following family of indifference curves between flexibility and postponement as shown in Figure 4.4. Table 4.2 summarizes the expected mismatch cost of all the 22 indifference curves as well as their optimality gap (as defined in Section 1.3.2) from the best curve (Curve #22).

| Curve # | Expected Cost | Optimality Gap | Curve # | Expected Cost | Optimality Gap |
|---|---|---|---|---|---|
| 1 | 4,815 | 504% | 12 | 993.5 | 24.53% |
| 2 | 4,345 | 445% | 13 | 889.3 | 11.47% |
| 3 | 3,945 | 394% | 14 | 837.5 | 4.98% |
| 4 | 3,605 | 352% | 15 | 827.1 | 3.67% |
| 5 | 3,100 | 289% | 16 | 818.5 | 2.59% |
| 6 | 2,550 | 220% | 17 | 809.5 | 1.47% |
| 7 | 2,440 | 206% | 18 | 804.6 | 0.85% |
| 8 | 1,905 | 139% | 19 | 801.3 | 0.44% |
| 9 | 1,480 | 86% | 20 | 799.1 | 0.16% |
| 10 | 1,215 | 52% | 21 | 798.3 | 0.06% |
| 11 | 1,050 | 32% | 22 | 797.8 | 0.00% |

*Tab. 4.2:* Mismatch Cost Values and Optimality Gaps for Flexibility-Postponement Indifference Curves

From Figure 4.4 and Table 4.2, we can see that for the dedicated system, even with full postponement, the performance gap is a shocking 206% (i.e. the cost is more than three times that of the first-best solution). As for

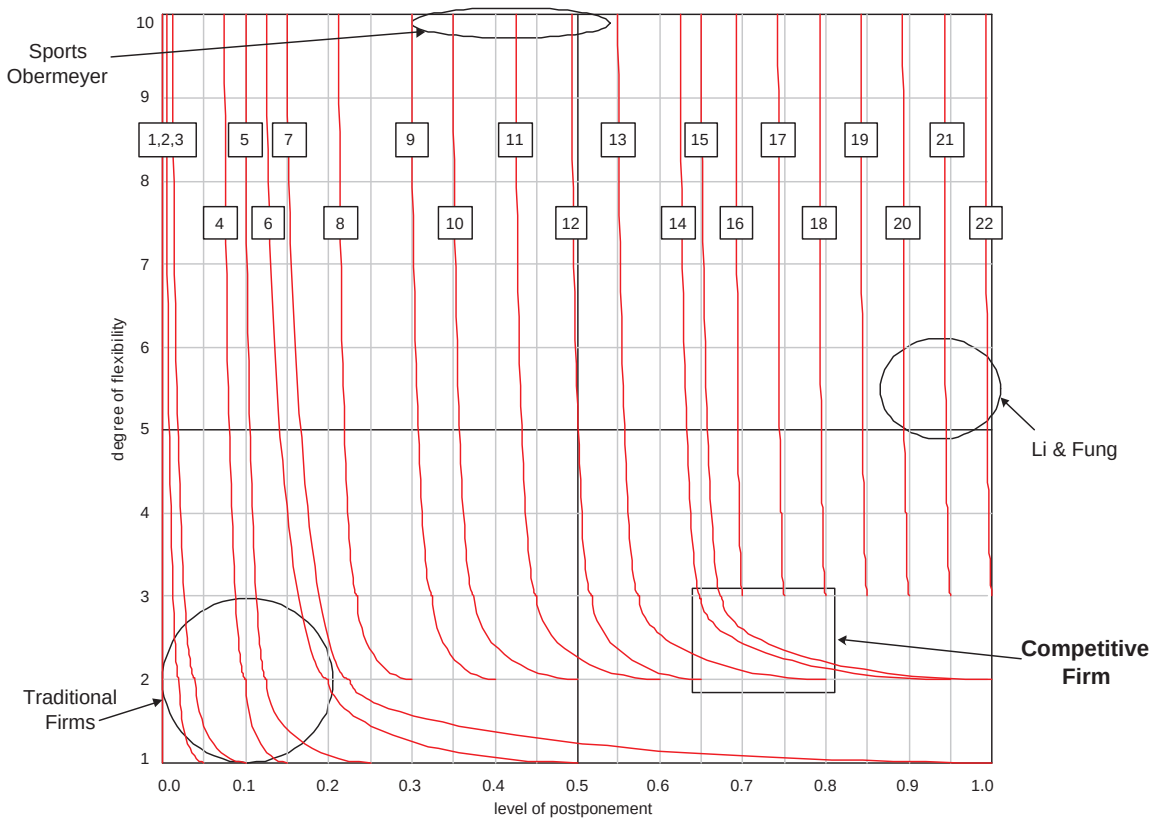


*Fig. 4.4:* Indifference Curves for Flexibility and Postponement

the 2-chain, postponement levels lower than 80% will bring the optimality gap to more than 11.5%. Fortunately, the 3-chain can still survive a 57% postponement level to provide a performance gap of only 11.5%. To further support the case of the 3-chain, any improvement to 4-chain and beyond is mostly marginal in the sense that such improvement can only replace a minimal amount of postponement.

We also try to identify the location of some firms on the indifference map to provide perspective on what one can achieve realistically. Traditional firms tend to be on the **lower left quadrant** and incurs ridiculously high expected costs. At the other end, the gold standard of manufacturing excellence Li & Fung lies somewhere along Curves #20, 21, or 22. Because of its clout, this company may very well be the only one of its kind that can afford both high levels of postponement (90% to 100%) and flexibility (5-chain, 6-chain or up), thus operating at almost optimality (**upper right quadrant**). Sport Obermeyer, studied by Fisher and Raman [23], probably falls somewhere in the **upper left quadrant** because it operates at full flexibility with some postponement. Its optimality gap is probably between 25% and 50%. These numbers suggest that the company may have overinvested in flexibility and could channel some of their resources to improve postponement.

Finally, we locate an area (**lower right quadrant**) where investments in flexibility and postponement appear to be most efficient for a firm without the power of Li & Fung. The first recommendation is to use a 3-chain with 70% postponement (Curve #16) which will result in an optimality loss of only 2.59%. The postponement level of this 3-chain system may be reduced to 65% (Curve #14) if one can tolerate a larger optimality loss of

about 5%. For an even lower budget, one may also consider a 3-chain with 57% postponement or a 2-chain with 80% postponement (Curve #13), both generating optimality losses of about 11.5%.

**As the System Grows**

The previous observations provide evidence of the potential value of a third chain. In particular, the third chain can be used to compensate for the loss brought about by partial postponement. This naturally begs the question of whether this theory will still hold when the system size grows. To this end, we reproduce the above numerical study to systems of size $n = 15, 20, 25, \ldots, 40$. We consider $d = 1, 2, 3, n$ which are the dedicated system, the 2-chain, the 3-chain, and full flexibility, respectively. As they are the recommended postponement levels in the previous discussion, we present the results for postponement levels of 65%, 70%, and 75% in Tables 4.3, 4.4, and 4.5, respectively.

| Size | Optimality Gap | | | | Ratios | |
|---|---|---|---|---|---|---|
| $n$ | $\mathcal{D}(n)$ | $\mathcal{C}_2(n)$ | $\mathcal{C}_3(n)$ | $\mathcal{F}(n)$ | $\frac{\mathcal{C}_2(n)-\mathcal{C}_3(n)}{\mathcal{C}_2(n)-\mathcal{F}(n)}$ | $\frac{\mathcal{C}_2(n)-\mathcal{C}_3(n)}{\mathcal{C}_2(n)}$ |
| 10 | 209.34% | 24.53% | 4.58% | 4.27% | 98.46% | 81.34% |
| 15 | 274.80% | 49.97% | 9.98% | 6.34% | 91.66% | 80.04% |
| 20 | 346.40% | 75.33% | 17.04% | 7.96% | 86.53% | 77.38% |
| 25 | 393.20% | 95.24% | 26.59% | 8.27% | 78.94% | 72.09% |
| 30 | 488.75% | 121.48% | 37.02% | 9.95% | 75.73% | 69.53% |
| 35 | 489.03% | 128.88% | 45.25% | 10.48% | 70.63% | 64.89% |
| 40 | 534.50% | 144.07% | 50.79% | 11.44% | 70.33% | 64.74% |

*Tab. 4.3:* Optimality Gap as Size Increases for 65% Postponement

One can clearly see that a 2-chain can perform extremely badly when system size grows. In fact, when $n = 40$ the optimality loss rises to as high as 144% at 65% postponement. This can be partially remedied by employing

| Size | Optimality Gap | | | | Ratios | |
|---|---|---|---|---|---|---|
| $n$ | $\mathcal{D}(n)$ | $\mathcal{C}_2(n)$ | $\mathcal{C}_3(n)$ | $\mathcal{F}(n)$ | $\frac{\mathcal{C}_2(n)-\mathcal{C}_3(n)}{\mathcal{C}_2(n)-\mathcal{F}(n)}$ | $\frac{\mathcal{C}_2(n)-\mathcal{C}_3(n)}{\mathcal{C}_2(n)}$ |
| 10 | 207.86% | 19.01% | 2.59% | 2.47% | 99.24% | 86.37% |
| 15 | 272.76% | 41.69% | 6.03% | 3.69% | 93.83% | 85.53% |
| 20 | 343.92% | 64.92% | 11.02% | 4.71% | 89.52% | 83.02% |
| 25 | 390.71% | 83.82% | 18.89% | 4.89% | 82.26% | 77.46% |
| 30 | 485.74% | 107.97% | 27.10% | 5.99% | 79.30% | 74.90% |
| 35 | 485.93% | 115.31% | 34.55% | 6.22% | 74.03% | 70.04% |
| 40 | 531.26% | 129.36% | 39.53% | 6.90% | 73.36% | 69.44% |

*Tab. 4.4:* Optimality Gap as Size Increases for 70% Postponement

| Size | Optimality Gap | | | | Ratios | |
|---|---|---|---|---|---|---|
| $n$ | $\mathcal{D}(n)$ | $\mathcal{C}_2(n)$ | $\mathcal{C}_3(n)$ | $\mathcal{F}(n)$ | $\frac{\mathcal{C}_2(n)-\mathcal{C}_3(n)}{\mathcal{C}_2(n)-\mathcal{F}(n)}$ | $\frac{\mathcal{C}_2(n)-\mathcal{C}_3(n)}{\mathcal{C}_2(n)}$ |
| 10 | 207.00% | 14.80% | 1.46% | 1.43% | 99.76% | 90.13% |
| 15 | 271.46% | 34.81% | 3.44% | 2.04% | 95.72% | 90.11% |
| 20 | 342.33% | 56.15% | 6.86% | 2.64% | 92.13% | 87.79% |
| 25 | 389.14% | 74.21% | 13.33% | 2.76% | 85.21% | 82.04% |
| 30 | 483.90% | 96.65% | 19.64% | 3.58% | 82.75% | 79.68% |
| 35 | 483.94% | 103.86% | 26.20% | 3.58% | 77.44% | 74.77% |
| 40 | 529.22% | 116.84% | 30.82% | 4.08% | 76.28% | 73.62% |

*Tab. 4.5:* Optimality Gap as Size Increases for 75% Postponement

the third chain, reducing the optimality loss by about 65% to 74% (e.g. from 144% to 51%). Now suppose you have already installed a 2-chain as recommended in existing literature, only to find out your postponement level is between 65% and 75%. Because of your poor performance, you consider upgrading to a 3-chain. Column 6 tells you that whatever benefit you may obtain from upgrading to full flexibility, you can already achieve at least 70% of it by just adding the third chain. Similarly, the last column informs us that whatever benefit you can get from upgrading to optimality (full flexibility and full postponement – very costly), you can already accrue at least about 65% of it by merely adding the third chain. These percentages will be even higher if system size is smaller than 40.

### *4.1.5 The Asymmetric Case*

In this section, we test our theory on the asymmetric setting. Here, the asymmetric analog of the 3-chain is a sparse structure with an additional layer of flexibility. For example, in an $n \times n$ system, the "2-chain" consists of $2n$ links while the "3-chain" is made up of $3n$ links. The resulting structures may no longer be regular and symmetric as in the earlier setting. Of course, the choice of which $2n$ or $3n$ links must be optimized. In the two numerical studies that follow, the constraint sampling methodology developed in [16] is utilized. The two case studies we consider are: (1) O'neill Inc. (in Cachon and Terwiesch [14]), and (2) Sport Obermeyer Ltd. (in Hammond and Anand [29]).

**O'neill Inc.**

O'neill Inc. is a designer and manufacturer of water sports apparel, wetsuits and accessories. Their products range for entry-level products for beginners, to wetsuits for competitive surfers, to sophisticated sportswear for professional divers. Because the demands for the products are subject to the whims of fashion, O'neill faces the tough production challenge of not ordering too much or too little. Table 4.6 presents a sample list of products and the respective demand forecasts (in terms of mean and standard deviation). In addition to its own production facility in Mexico, O'neill also employs the services of contract manufacturers in Asia. The issue of which facility to be equipped with technology to produce which product is basically the flexibility question we are interested in. Moreover, a portion of production must be committed before the selling season, while the rest can be allocated later on with better idea of the actual demand.

| Item | Model | Expected demand | Standard deviation |
|---|---|---|---|
| 1 | DIVE COMP 3/2 FULL | 1,100 | 660 |
| 2 | WMS 7000 X 7MM FULL | 600 | 360 |
| 3 | EPIC 5/3 W/HD | 800 | 296 |
| 4 | HEAT 3/2 | 1,200 | 444 |
| 5 | HEATWAVE 4/3 | 700 | 259 |
| 6 | ZEN-ZIP 4/3 | 3,100 | 1,147 |
| 7 | TRIATHLON 4/3 FULL | 2,600 | 1,690 |
| 8 | REACTOR 3/2 | 1,500 | 750 |
| 9 | CYCLONE 4/3 | 950 | 665 |
| 10 | WMS EVOLUTION 4/3 | 850 | 595 |

*Tab. 4.6:* Demand Forecasts for Diving Products at O'neill Inc.

In this study, we consider a production network of 10 facilities, each

one primarily assigned to manufacture one product. For example, facility 1 mainly produces DIVE COMP 3/2 FULL. Also, each facility has enough capacity to meet the expected demand of its primary product, e.g. facility 1 has a capacity of 1,100 units. To make the system more flexible, we add 10 more links and call the resulting 20-link structure *2-sparse*, the asymmetric analog of the 2-chain. Augmenting with another set of 10 links, we would get a *3-sparse* structure, the asymmetric analog of the 3-chain. We then examine the performance of the 2-sparse and the 3-sparse structures relative to full flexibility, under varying levels of production postponement.

Next, we briefly explain how the numerical study is carried out. Here, $\mathbf{D} = (D_1, D_2, \ldots, D_{10})$ is the demand vector while $\mathbf{C} = (C_1, C_2, \ldots, C_{10})$ is the capacity vector.

**Numerical Procedure:**

**[Step 1]**: Estimate the sampling probabilities for each facility-product pair $(i, j)$ such that $i \neq j$. This represents the likelihood that link $(i, j)$ will be the next most attractive link to add to the dedicated system (cf. Chou et al. [16] for details on the constraint sampling methodology).

1. Generate 1000 demand realizations based on the given demand distribution. Let $\mathbf{D}^k$ denote the demand generated in the $k$th instance.

2. Estimate the weight assigned to link $(i, j)$, using

$$w_{ij} = \frac{1}{1000} \sum_{k=1}^{1000} \frac{(D_i^k - C_i)^+ (C_j - D_j^k)^+}{\max\{\sum_{l=1}^{10} (D_l^k - C_l)^+, \sum_{l=1}^{10} (C_l - D_l^k)^+\}}$$

3. Estimate the sampling probability for link $(i, j)$, using

$$\tilde{p}_{ij} = \frac{w_{ij}}{\sum_{k=1}^{10} \sum_{l=1}^{10} w_{kl}}$$

[**Step 2**]: Generate the best 2-sparse structure. Call it $\mathcal{S}_2(10)$.

1. Generate 100 2-sparse structures. For each 2-sparse structure, sample one link at a time based on the sampling probabilities $\tilde{p}_{ij}$, and add to the dedicated system until 10 new links have been added.

2. Grade the 100 2-sparse structures and choose the best. Generate another 1000 demand realizations, and simulate the performance of each 2-sparse structure based on this set of demand. Select the 2-sparse structure with the best performance.

[**Step 3**]: Generate the best 3-sparse structure. Call it $\mathcal{S}_3(10)$.

1. Generate 100 3-sparse structures. For each 3-sparse structure, sample one link at a time based on the sampling probabilities $\tilde{p}_{ij}$ excluding those already included in $\mathcal{S}_2(10)$. Add the link to the best 2-sparse structure and repeat until 10 new links have been added.

2. Grade the 100 3-sparse structures and choose the best. Use the 1000 demand realizations generated for 2-sparse performance evaluation, and simulate the performance of each 3-sparse structure based on this set of demand. Select the 3-sparse structure with the best performance.

[**Step 4**]: Compute the performance of the dedicated system, the 2-sparse system, the 3-sparse system, and full flexibility.

1. Generate 100 demand scenarios. Let $\mathbf{D}^k$ denote the demand generated in the $k$th scenario.

2. For each system $\mathcal{G}(10) \in \{\mathcal{D}(10), \mathcal{S}_2(10), \mathcal{S}_3(10), \mathcal{F}(10)\}$ and for each postponement level $\alpha \in \{0.00, 0.05, 0.10, \ldots, 0.95, 1.00\}$, solve the following large LP as an approximation of the underlying stochastic newsvendor problem. For simplicity, we assume equal overage and underage costs. The decision vectors $\mathbf{x}$ and $\mathbf{y}^k$ denote first-stage production and second-stage production, respectively. For given structure $\mathcal{G}(10)$ and postponement level $\alpha$, the minimum expected mismatch cost is as follows.

$$
\begin{aligned}
& G^*_{\mathcal{G}(10)}(\alpha) \\
& \quad = \min \sum_{k=1}^{100} \sum_{i=1}^{10} \left[ \left( \sum_{j=1}^{10} (x_{ij} + y_{ij}^k) - D_i^k \right)^+ + \left( D_i^k - \sum_{j=1}^{10} (x_{ij} + y_{ij}^k) \right)^+ \right] \\
& \qquad \text{s.t.} \sum_{i:(i,j)\in\mathcal{G}(10)} x_{ij} \le (1-\alpha) C_j, \quad \forall j = 1, \ldots, 10 \\
& \qquad\quad \sum_{i:(i,j)\in\mathcal{G}(10)} y_{ij}^k \le \alpha C_j, \quad \forall j = 1, \ldots, 10, \ \forall k = 1, \ldots, 100 \\
& \qquad\quad x_{ij} \ge 0, \quad \forall\ i, j = 1, \ldots, n \\
& \qquad\quad x_{ij} = 0, \quad \forall\ (i,j) \notin \mathcal{G}(10) \\
& \qquad\quad y_{ij}^k \ge 0, \quad \forall\ i, j = 1, \ldots, n, \ \forall k = 1, \ldots, 100 \\
& \qquad\quad y_{ij}^k = 0, \quad \forall\ (i,j) \notin \mathcal{G}(10), \ \forall k = 1, \ldots, 100.
\end{aligned}
$$

3. For $\mathcal{S}_2(10)$ and $\mathcal{S}_3(10)$, compute the flexibility efficiency as follows.

$$FE(\mathcal{G}(10), \alpha) = \frac{G^*_{\mathcal{D}(10)}(\alpha) - G^*_{\mathcal{G}(10)}(\alpha)}{G^*_{\mathcal{D}(10)}(\alpha) - G^*_{\mathcal{F}(10)}(\alpha)}$$

| | Expected Mismatch Cost | | | | Flex. efficiency | |
|---|---|---|---|---|---|---|
| $\alpha$ | $\mathcal{D}(10)$ | $\mathcal{S}_2(10)$ | $\mathcal{S}_3(10)$ | $\mathcal{F}(10)$ | $\mathcal{S}_2(10)$ | $\mathcal{S}_3(10)$ |
| 0.00 | 4,844 | 4,824 | 4,815 | 4,815 | 67.24% | 100.00% |
| 0.05 | 4,545 | 4,360 | 4,277 | 4,211 | 55.29% | 80.26% |
| 0.10 | 4,274 | 3,939 | 3,798 | 3,685 | 56.94% | 80.88% |
| 0.15 | 4,020 | 3,560 | 3,370 | 3,230 | 58.21% | 82.28% |
| 0.20 | 3,790 | 3,221 | 2,983 | 2,826 | 58.98% | 83.73% |
| 0.25 | 3,587 | 2,923 | 2,648 | 2,483 | 60.10% | 85.05% |
| 0.30 | 3,406 | 2,659 | 2,349 | 2,189 | 61.36% | 86.83% |
| 0.35 | 3,251 | 2,432 | 2,088 | 1,937 | 62.31% | 88.46% |
| 0.40 | 3,121 | 2,236 | 1,869 | 1,727 | 63.47% | 89.76% |
| 0.45 | 3,009 | 2,070 | 1,688 | 1,558 | 64.68% | 91.00% |
| 0.50 | 2,915 | 1,931 | 1,538 | 1,428 | 66.20% | 92.62% |
| 0.55 | 2,837 | 1,816 | 1,415 | 1,324 | 67.50% | 94.01% |
| 0.60 | 2,773 | 1,717 | 1,312 | 1,237 | 68.75% | 95.16% |
| 0.65 | 2,719 | 1,634 | 1,230 | 1,169 | 69.99% | 96.10% |
| 0.70 | 2,676 | 1,575 | 1,169 | 1,124 | 70.93% | 97.09% |
| 0.75 | 2,647 | 1,533 | 1,125 | 1,095 | 71.79% | 98.05% |
| 0.80 | 2,623 | 1,503 | 1,095 | 1,071 | 72.21% | 98.48% |
| 0.85 | 2,605 | 1,479 | 1,071 | 1,054 | 72.58% | 98.91% |
| 0.90 | 2,589 | 1,463 | 1,060 | 1,047 | 73.00% | 99.15% |
| 0.95 | 2,580 | 1,451 | 1,053 | 1,041 | 73.36% | 99.25% |
| 1.00 | 2,577 | 1,447 | 1,051 | 1,041 | 73.61% | 99.38% |

*Tab. 4.7:* Expected Mismatch Cost and Flexibility Efficiency for O'neill Inc.

Table 4.7 summarizes the result for the numerical study of O'neill Inc. The general behavior of the production system under different levels of flexibility and postponement appears to be similar to the symmetric case. When there is full postponement, the 2-sparse structure performs quite well at 73%

efficiency. However, under partial postponement, the efficiency can drop to the range 55-65%. By adding a layer of flexibility, the 3-sparse structure not only recovers the said flexibility loss, but also provides an efficiency of at least 80% under all postponement levels. This is even better than the 73% of the 2-sparse structure under full postponement.

**Sport Obermeyer Ltd.**

Sport Obermeyer Ltd. is a manufacturer of stylish high-performance ski clothing and ski equipment products. It holds a commanding 45% share of the children's skiwear market and 11% share of the adult market. One of its main challenges is the production planning for women's parkas mostly due to the demand volatility caused by ever changing fashion. Table 4.8 shows ten styles of women's parkas and their respective demand forecasts. A clearer picture of actual demand can be obtained through feedback from retailers, although such information will only be available after the Las Vegas trade show in March of each year. However, manufacturing has to start well in advance of the trade show season and some items must be produced to stock. Fortunately, a considerable portion of capacity (called reactive capacity) may be deferred until March. The level of postponement (i.e. the amount of reactive capacity) and the degree of flexibility in Sport Obermeyer's production network play huge roles in determining the outcome of the company's financial performance.

In this study, we consider a production network of 10 facilities, each one primarily assigned to manufacture one product. For example, facility 1 mainly produces the Gail style. Also, each facility has enough capacity to meet the expected demand of its primary product, e.g. facility 1 has a capac-

| Item | Style | Expected demand | Standard deviation |
|---|---|---|---|
| 1 | Gail | 1,017 | 194 |
| 2 | Isis | 1,042 | 323 |
| 3 | Entice | 1,358 | 248 |
| 4 | Assault | 2,525 | 340 |
| 5 | Teri | 1,100 | 381 |
| 6 | Electra | 2,150 | 404 |
| 7 | Stephanie | 1,113 | 524 |
| 8 | Seduced | 4,017 | 556 |
| 9 | Anita | 3,296 | 1,047 |
| 10 | Daphne | 2,383 | 697 |

*Tab. 4.8:* Demand Forecasts for Women's Parkas at Sport Obermeyer

ity of 1,017 units. Although facilities have their primary style assignments, it would serve the company well if these facilities can also produce the other styles. As much as full flexibility whereby all facilities can make all styles is most desirable, we show that a 3-sparse structure with only 30 links (20 links in addition to the primary assignments) already captures a large portion of the benefits of full flexibility. Similar to the symmetric case as well as the case study on O'neill Inc., this finding is true for all levels of production postponement. In fact, for most levels of postponement, the 3-sparse structure is already at least 90% efficient. We use the same numerical procedure as in the case on O'neill Inc. Table 4.9 summarizes the results.

## *4.2 Process Flexibility and Supply Disruptions*

In this section, we further discuss the merits of the third chain through the lens of supply disruptions. Recent studies have pointed out that supply chains

| | Expected Mismatch Cost | | | | Flex. efficiency | |
|---|---|---|---|---|---|---|
| $\alpha$ | $\mathcal{D}(n)$ | $\mathcal{S}_2(n)$ | $\mathcal{S}_3(n)$ | $\mathcal{F}(n)$ | $\mathcal{S}_2(n)$ | $\mathcal{S}_3(n)$ |
| 0.00 | 3,500 | 3,480 | 3,477 | 3,477 | 88.16% | 100.00% |
| 0.05 | 3,066 | 2,809 | 2,681 | 2,574 | 52.22% | 78.37% |
| 0.10 | 2,719 | 2,311 | 2,059 | 1,918 | 50.94% | 82.39% |
| 0.15 | 2,463 | 1,939 | 1,602 | 1,465 | 52.50% | 86.32% |
| 0.20 | 2,268 | 1,664 | 1,270 | 1,158 | 54.37% | 89.84% |
| 0.25 | 2,127 | 1,463 | 1,038 | 952 | 56.51% | 92.71% |
| 0.30 | 2,023 | 1,316 | 878 | 809 | 58.24% | 94.27% |
| 0.35 | 1,945 | 1,208 | 771 | 707 | 59.52% | 94.84% |
| 0.40 | 1,886 | 1,127 | 696 | 645 | 61.20% | 95.88% |
| 0.45 | 1,843 | 1,065 | 646 | 610 | 63.15% | 97.09% |
| 0.50 | 1,813 | 1,022 | 617 | 591 | 64.71% | 97.88% |
| 0.55 | 1,793 | 990 | 607 | 584 | 66.41% | 98.11% |
| 0.60 | 1,780 | 967 | 600 | 584 | 67.96% | 98.66% |
| 0.65 | 1,770 | 951 | 594 | 584 | 69.03% | 99.11% |
| 0.70 | 1,764 | 941 | 591 | 584 | 69.72% | 99.42% |
| 0.75 | 1,759 | 934 | 589 | 584 | 70.21% | 99.56% |
| 0.80 | 1,757 | 929 | 589 | 584 | 70.57% | 99.56% |
| 0.85 | 1,754 | 927 | 589 | 584 | 70.69% | 99.56% |
| 0.90 | 1,753 | 926 | 589 | 584 | 70.73% | 99.56% |
| 0.95 | 1,753 | 926 | 589 | 584 | 70.75% | 99.56% |
| 1.00 | 1,752 | 926 | 589 | 584 | 70.74% | 99.56% |

*Tab. 4.9:* Expected Mismatch Cost and Flexibility Efficiency for Sport Obermeyer

are increasingly susceptible to disruptions that may be caused by labor strikes (e.g. General Motors [12]), hurricanes (e.g. Katrina and Rita [44]), fires (e.g. Philips semiconductor plant [36]), and other unexpected calamities. Given the millions of dollars lost in these mishaps, there has been growing concern and interest in the study of supply disruption mitigation. One important finding in the literature is that supply disruption risks are fundamentally distinct from demand uncertainty and recurrent supply risks, and thus may require a new set of strategies (cf. [50], [15]). For example, Chopra et al. [15] show the importance of decoupling recurrent supply risk and disruption risk when planning mitigation strategies, while Snyder and Shen [50] demonstrate that the optimal strategy for coping with supply disruptions is the exact opposite of that for demand uncertainty.

Without consideration for supply disruptions, the thought leaders in recent decades have championed tightly optimized and just-in-time practices. Unfortunately, the increasing threat of supply disruptions has exposed the vulnerability of these lean supply chain systems. Therefore, recent research makes a case for more redundancy or slack in order to buffer against disruption uncertainty. However, firms have historically been disinclined to invest in additional infrastructure or inventory, despite the potentially large payoff in the event of a disruption. Hence, it is but natural to turn to process flexibility for a way to reduce the buffer requirements or to maximize the utilization of additional resources.

Tomlin and Wang [52] examine both mix flexibility and dual sourcing in an attempt to study supply chains characterized by both demand uncertainty and unreliable resources. For a firm that can invest in dedicated resources and

fully flexible resources, they show that the intuition that a flexible strategy dominates a dedicated strategy is true only if the firm is risk neutral or if the resources are perfectly reliable. When both conditions fail to hold, a resource-aggregation disadvantage inherent in their model may outweigh the capacity pooling benefits of flexibility. However, this paper does not address the issue of partial flexibility.

A more related paper is by Lim et al. [40] who extend the classical Jordan and Graves [32] study on process flexibility with the possibility of supply disruptions and the underlying cost of flexibility. They focus on single failures (i.e. single link disruptions and single node disruptions) because the scenario of multiple failures can be decomposed into subnetworks with single failures for the purpose of analysis. A measure called "fragility" is introduced which quantifies the change in system performance before and after a disruption. Under some conditions, it can be shown that reducing system fragility is equivalent to reducing total cost of the system. That said, their main result is as follows: if a system is more exposed to link disruptions, it is preferable to install a collection of small chains, while it is preferable to build a longer chain when the network is more exposed to node disruptions.

### *4.2.1 Fragility and Flexibility*

In this section, we want to examine how the third chain (or additional flexibility) can be utilized to improve the fragility of the system. To this end, we first recall the definition of fragility (cf. [40]).

**Definition 3.** *The fragility of a system $\mathcal{G}(n)$ with respect to disruption $D$ is*

*the difference in expected shortfalls after and before the disruption. That is,*

$$F^D(\mathcal{G}(n)) = SF^D(\mathcal{G}(n)) - SF(\mathcal{G}(n)).$$

Because expected shortfall is expected total demand minus expected total system sales, system fragility can alternatively be expressed as the difference in expected total system sales before and after the disruption, i.e.

$$F^D(\mathcal{G}(n)) = \mathrm{E}[Z^*_{\mathcal{G}(n)}(\mathbf{D})] - \mathrm{E}[Z^{*D}_{\mathcal{G}(n)}(\mathbf{D})] \tag{4.1}$$

where $Z^*_{\mathcal{G}(n)}(\mathbf{D})$ and $Z^{*D}_{\mathcal{G}(n)}(\mathbf{D})$ are the respective maximum flows before and after the disruption $D$ given demand realization $\mathbf{D}$, as defined in Section 1.3.1.

In Lim et al. [40], the rationale for focusing on system fragility when comparing two or more systems is the assumption that these systems have equal (or comparable) total system costs when there is no disruption. For example, compare a long 2-chain in a $10 \times 10$ production network with a collection of 5 short 2-chains. Each of the latter is just a fully flexible $2 \times 2$ system. Suppose that the additional cost of installing the long 2-chain is just about recovered by the increased expected sales (equivalently, reduced shortfall) in the system. Lim et al. [40] show that short chains have lower fragility than the long chain under single link disruptions, while the opposite holds under single node disruptions.

In a similar breath, we suppose that the 3-chain has already been ratio-

nalized on the basis of increased system size and/or the presence of partial production postponement as in previous sections. That is, we assume that the expected total system cost of the 3-chain does not exceed that of the 2-chain when there is no disruption. We then examine the fragility of the 3-chain vis-à-vis the 2-chain following equation (4.1). For our simulation study, we consider symmetric systems of size $n = 5, 10, \ldots, 40$. Capacity at each facility is set at 2000 units while demand is normally distributed with mean $\mu = 2000$, truncated from 0 to $2\mu$, and coefficient of variation $CV = 0.2, 0.3, 0.4, 0.5$. We simulate 1000 demand scenarios for each combination and summarize the results in Table 4.10. For all system sizes, all CV values, and both disruption types, we find that the 3-chain is significantly less fragile than the 2-chain. That is,

$$F^D(\mathcal{C}_3(n)) < F^D(\mathcal{C}_2(n))$$

$\forall D \in \{1LD, 1ND\}, n \in \{5, 10, \ldots, 40\}, CV \in \{0.2, 0.3, 0.4, 0.5\}$. This serves as additional evidence for the value of the third chain on top of the arguments already established in Section 2.4.4 and Section 4.1. However, there is no consistent pattern on the behavior of fragility with respect to system size or demand coefficient of variation.

We also conduct a comparative study between a long 3-chain and a collection of short 3-chains (fully flexible $3 \times 3$ systems). For system size $n = 6, 12, \ldots, 30$, we consider the same demand, supply, and disruption scenarios as in the previous study. Table 4.11 compiles the results. We observe that Lim et al.'s [40] result that short 2-chains are better under link disrup-

| Size $n$ | Disrupt Type | Coefficient of Variation | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 20% | | 30% | | 40% | | 50% | |
| | | $\mathcal{C}_2(n)$ | $\mathcal{C}_3(n)$ | $\mathcal{C}_2(n)$ | $\mathcal{C}_3(n)$ | $\mathcal{C}_2(n)$ | $\mathcal{C}_3(n)$ | $\mathcal{C}_2(n)$ | $\mathcal{C}_3(n)$ |
| 5 | Link | 165 | 0 | 230 | 0 | 292 | 0 | 337 | 4 |
| | Node | 1677 | 1,676 | 1,498 | 1,481 | 1,474 | 1,418 | 1,469 | 1,361 |
| 10 | Link | 282 | 0 | 417 | 2 | 524 | 17 | 475 | 25 |
| | Node | 1,549 | 1,519 | 1,510 | 1,394 | 1,488 | 1,281 | 1,535 | 1,312 |
| 15 | Link | 419 | 1 | 529 | 27 | 563 | 59 | 562 | 80 |
| | Node | 1,555 | 1,486 | 1,567 | 1,360 | 1,559 | 1,304 | 1,561 | 1,269 |
| 20 | Link | 437 | 5 | 571 | 40 | 560 | 93 | 562 | 138 |
| | Node | 1,514 | 1,380 | 1,594 | 1,333 | 1,551 | 1,293 | 1,583 | 1,352 |
| 25 | Link | 540 | 11 | 558 | 70 | 582 | 125 | 562 | 177 |
| | Node | 1,579 | 1,385 | 1,565 | 1,279 | 1,585 | 1,351 | 1,569 | 1,331 |
| 30 | Link | 566 | 19 | 585 | 97 | 577 | 143 | 579 | 202 |
| | Node | 1,596 | 1,365 | 1,608 | 1,337 | 1,544 | 1,330 | 1,549 | 1,344 |
| 35 | Link | 577 | 32 | 584 | 105 | 540 | 147 | 547 | 203 |
| | Node | 1,583 | 1,333 | 1,600 | 1,372 | 1,590 | 1,400 | 1,522 | 1,393 |
| 40 | Link | 626 | 42 | 570 | 116 | 588 | 191 | 511 | 183 |
| | Node | 1,598 | 1,327 | 1,553 | 1,313 | 1,559 | 1,361 | 1,533 | 1,387 |

*Tab. 4.10:* Fragility for 2-Chain and 3-Chain under Single Link and Single Node Disruptions for Various Levels of Demand Uncertainty

tions while a long 2-chain is better under node disruptions extends nicely to the case of 3-chains. In fact, the short 3-chains under link disruptions are not fragile at all in all the situations we considered. Also, for small to medium-sized production networks ($n = 6, \ldots, 24$) and coefficient of variation not too large (at most 40%), the long 3-chain is not too shabby under link disruptions while the short 3-chains can perform quite poorly under node disruptions. Having said that, we would tend to recommend a long 3-chain over a collection of short 3-chains for most realistic cases whereby the nature of the next disruption (link or node) is unknown.

| Size | Disrupt | Coefficient of Variation | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 20% | | 30% | | 40% | | 50% | |
| $n$ | Type | long | short | long | short | long | short | long | short |
| 6 | Link | 0 | 0 | 0 | 0 | 1 | 0 | 7 | 0 |
| | Node | 1,614 | 1,721 | 1,478 | 1,581 | 1,391 | 1,518 | 1,393 | 1,485 |
| 12 | Link | 1 | 0 | 9 | 0 | 40 | 0 | 63 | 0 |
| | Node | 1,489 | 1,726 | 1,364 | 1,600 | 1,321 | 1,553 | 1,331 | 1,455 |
| 18 | Link | 1 | 0 | 26 | 0 | 78 | 0 | 126 | 0 |
| | Node | 1,447 | 1,727 | 1,322 | 1,602 | 1,277 | 1,468 | 1,340 | 1,465 |
| 24 | Link | 8 | 0 | 67 | 0 | 113 | 0 | 164 | 0 |
| | Node | 1,348 | 1,712 | 1,326 | 1,584 | 1,313 | 1,518 | 1,340 | 1,485 |
| 30 | Link | 17 | 0 | 81 | 0 | 138 | 0 | 189 | 0 |
| | Node | 1,383 | 1,739 | 1,284 | 1,582 | 1,346 | 1,485 | 1,344 | 1,461 |

*Tab. 4.11:* Fragility for Long 3-Chain versus Short 3-Chain under Single Link and Single Node Disruptions for Various Levels of Demand Uncertainty

### *4.2.2 Fragility, Flexibility and Capacity*

We have already seen that supply disruptions (particularly, node disruptions) can lead to significant losses to the system as exhibited by the fragility values in Table 4.10. Although the third chain can help mitigate these losses to some

extent, we do not expect the same benefits to come from further upgrades in flexibility. Hence, we turn to increasing capacity as another mitigation strategy. Suppose we are given a budget to increase capacity by 10%. How will this affect system fragility? More importantly, how must this extra capacity be utilized? For example, is it more desirable to distribute the additional capacity evenly among the existing facilities or to place that extra capacity in a standby facility that will fill in for whichever facility gets disrupted?

We carry out a simulation study to compare the two ways we can add capacity to systems exposed to supply disruptions. In the event of a disruption, the system incurs a total performance loss, which we break down into disruption loss and flexibility loss. Disruption loss is that portion that is due to the occurrence of the disruption, while flexibility loss is that which results from having partial flexibility instead of full flexibility. In our study, we focus on the flexibility loss because with this amount minimized, further measures to reduce disruption losses can take comfort in the fact that the system already operates at close to full flexibility. We consider system sizes $n = 5, 10, \ldots, 40$, and supply of 2000 units at each facility with additional 2000 units to be allocated either evenly among the existing facilities or housed in a standby facility. Demand is normally distributed with mean $\mu = 2000$, truncated from 0 to $2\mu$, and coefficient of variation is 0.3. We simulate 1000 demand scenarios and compute the flexibility efficiency as presented in Table 4.12. We observe that allocating the extra capacity to a standby facility appears to be a more robust approach when considering a 2-chain system. However, when a 3-chain is employed, the performance turns out to be insensitive to how the additional capacity is allocated. This provides the firm with

more decision flexibility, especially in cases when either of the two allocation options is not available. Ultimately, this implementation flexibility also adds to the value of having a third chain.

| Size $n$ | Disrupt Type | Use of 10% Capacity | | | |
|---|---|---|---|---|---|
| | | Distribute Evenly | | Standby Facility | |
| | | $\mathcal{C}_2(n)$ | $\mathcal{C}_3(n)$ | $\mathcal{C}_2(n)$ | $\mathcal{C}_3(n)$ |
| 5 | None | 100% | 100% | 100% | 100% |
| | Link | 94% | 100% | 100% | 100% |
| | Node | 98% | 100% | 100% | 100% |
| 10 | None | 99% | 100% | 98% | 100% |
| | Link | 90% | 100% | 98% | 100% |
| | Node | 91% | 100% | 98% | 100% |
| 15 | None | 97% | 100% | 95% | 100% |
| | Link | 87% | 100% | 95% | 100% |
| | Node | 86% | 100% | 95% | 100% |
| 20 | None | 93% | 100% | 92% | 100% |
| | Link | 86% | 99% | 92% | 100% |
| | Node | 84% | 99% | 92% | 100% |
| 25 | None | 92% | 100% | 90% | 100% |
| | Link | 86% | 99% | 90% | 100% |
| | Node | 84% | 99% | 90% | 100% |
| 30 | None | 90% | 100% | 88% | 99% |
| | Link | 84% | 99% | 88% | 99% |
| | Node | 82% | 98% | 88% | 99% |
| 35 | None | 87% | 99% | 86% | 99% |
| | Link | 83% | 98% | 86% | 99% |
| | Node | 81% | 97% | 86% | 99% |
| 40 | None | 87% | 99% | 86% | 98% |
| | Link | 84% | 98% | 86% | 98% |
| | Node | 82% | 97% | 86% | 98% |

*Tab. 4.12:* Flexibility Efficiency for Two Ways to Add Capacity to Symmetric Systems Exposed to Supply Disruptions

### 4.2.3 The Asymmetric Case

To test our observations on the asymmetric setting, we recall the two case studies considered in Section 4.1.5; namely, O'neill Inc. and Sport Obermeyer Ltd. Using the constraint sampling methodology introduced by Chou et al. [16], we generate the 2-sparse and 3-sparse structures for each case study. For each study, we simulate 1000 demand scenarios and compute the fragility values for both structures, and both single link and single node disruptions. Because the system is no longer symmetric, the expected system performance depends on which link or node is disrupted. We compute the fragility for each disruption and summarize the results in Figure 4.5. Although a very small set of instances shows that the 2-chain may even have lower fragility than the 3-chain, the 3-chain is for the most part significantly less fragile than the 2-chain. This supports our earlier finding that on top of reducing the negative effects of increasing system size and partial production postponement, the third chain can likewise improve system fragility in the event of unexpected supply disruptions.

We also examine whether it is more desirable to distribute an additional 10% capacity proportionately among existing facilities or place it in a standby facility. In Table 4.13, we find that the standby facility in a 2-sparse structure is more robust, especially when a node is disrupted. With the 3-sparse structure, the same is still true but the difference is no longer as pronounced. This supports our earlier finding that the performance of the 3-sparse structure is insensitive to the allocation method for extra capacity, giving the firm more implementation flexibility.

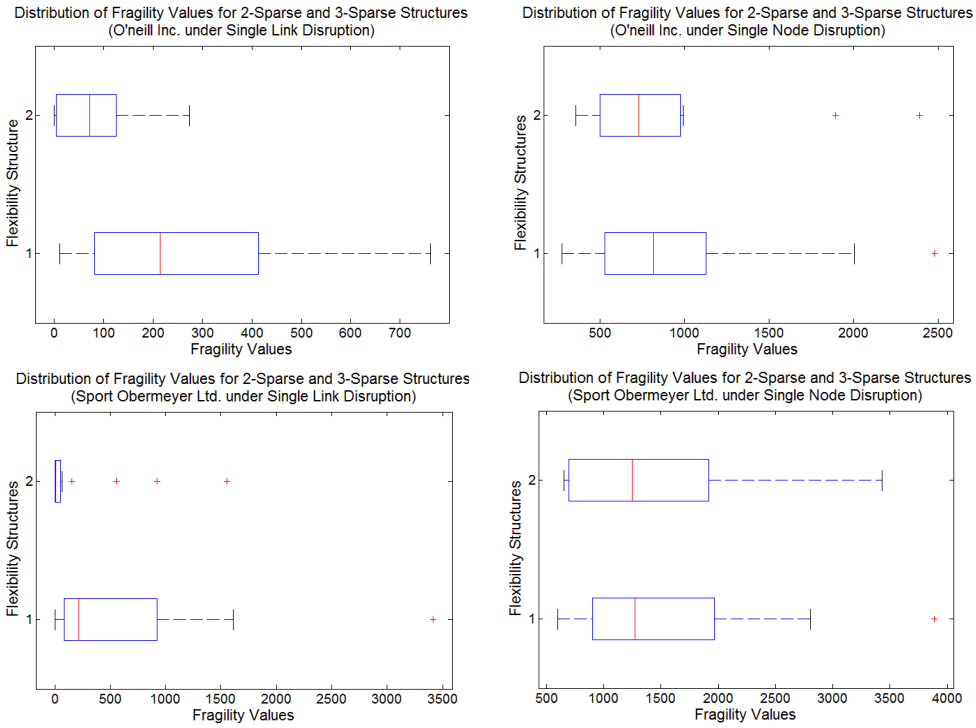


*Fig. 4.5:* Box and Whisker Plots for Fragility Values of 2-Sparse and 3-Sparse Structures of Asymmetric Systems Under Link and Node Disruptions

| Company Name | Disrupt Type | Use of 10% Capacity | | | |
|---|---|---|---|---|---|
| | | Distribute | | Standby Facility | |
| | | $\mathcal{S}_2(n)$ | $\mathcal{S}_3(n)$ | $\mathcal{S}_2(n)$ | $\mathcal{S}_3(n)$ |
| O'neill | None | 79% | 100% | 76% | 99% |
| | Link | 78% | 97% | 76% | 99% |
| | Node | 72% | 96% | 76% | 99% |
| Sport Ober-meyer | None | 75% | 99% | 70% | 99% |
| | Link | 71% | 95% | 70% | 99% |
| | Node | 64% | 95% | 70% | 99% |

*Tab. 4.13:* Flexibility Efficiency for Two Ways to Add Capacity to Asymmetric Systems Exposed to Supply Disruptions

# 5. CONCLUSIONS

The purpose of this thesis is to provide an analytical justification of why partial flexibility performs nearly as well as full flexibility, and to extend this theory of partial flexibility to environments that take into account other factors relevant to the practice of process flexibility or capacity pooling.

We first study the asymptotic performance of the chaining strategy when system size grows very large. For the symmetric case where supply and (mean) demand are balanced and identical, we develop a generalized random walk approach that can analytically compute the efficiency of chaining under general demand distributions. For uniform and normal demand distributions, the results show that the 2-chain already accrues at least 58% and 70%, respectively, of the benefits of full flexibility. This confirms the widely believed maxim in the community that chaining already accounts for most of the gains of full flexibility. Our method can also be adjusted to measure the performance of higher order chains, such as the 3-chain, the 4-chain, and so on.

Subsequently, we expand our analysis to take into account factors such as the response dimension, partial production postponement, and the occurrence of supply disruptions. In each scenario, we find that the performance of chaining may deteriorate significantly. We then propose measures on how

to reduce this performance decline.

When the response dimension is not perfect, we demonstrate that the performance of any flexible system may be seriously compromised. For example, when system response is sufficiently low, the chaining efficiency for a $10 \times 10$ system can go down to as low as 42.83%. This can be interpreted as a precaution not to overstate the benefits of process flexibility and as a call to examine the system response when engaging in process flexibility. Nevertheless, we find that surprisingly, when system response deteriorates to a certain threshold, the performance plateaus at a certain level and further response deterioration will cause no more harm to the system than it does to full flexibility. In addition, the performance of a long chain becomes identical to short chains, which differs from the high response case. This suggests that when system response is low and can no longer be improved, one can be better served by installing the less expensive shorter chains. We also show that given limited resources, upgrading system response outperforms upgrading system range in most cases. Moreover, improving system response can provide even more benefits when coupled with initiatives to reduce demand variability.

When full production postponement is not possible, we discover that previous results on partial flexibility no longer holds as strongly. In the example of small systems, we find that when postponement level is lower than 80%, the celebrated 2-chain strategy may perform quite badly, with a performance loss of more than 12%. By adding another layer of flexibility, i.e. a third chain, we find that the optimality loss improves to 5% even when postponement drops to 65%. For larger systems, we find that the

performance of the 2-chain becomes even worse, but the 3-chain under 65% to 75% postponement may be able to salvage a substantial portion of this optimality loss. Further flexibility upgrades, e.g. fourth or fifth chain, can no longer produce as much benefit. We also study the flexibility-postponement tradeoff and find that a firm operating with a 3-chain at 70% postponement can perform extremely well with minimal optimality loss.

Under the threat of supply disruptions, we find that the fragility of a 2-chain (both long and short) may be too high under both link and node disruptions. By introducing a third chain, the fragility of the system is significantly reduced. This suggests that in addition to cushioning the adverse effects of system size increase and partial production postponement, a third chain can also provide some protection against supply disruptions. Because redundancy is another widely recommended strategy for supply risk mitigation, we also study the interaction of flexibility and additional capacity. We observe that when using a 3-chain, the choice of how to allocate additional capacity no longer becomes critical, which differs from the case when a 2-chain is used. This provides implementation flexibility for the decision-maker, further strengthening the case for the value of the third chain. Although it can be argued that the above benefits can also be obtained in a 4-chain or higher chains, one must bear in mind that it is in the 3-chain that these benefits first appear and additional benefits must be established for additional chains.

There are several other directions to further extend the results in this thesis. It will be interesting to consider price-responsive demands and formulate the manufacturers problem as one of maximizing profits. It would also be interesting to look at this problem in an oligopolistic framework and

to examine the impact of pricing and partial flexibility on the strategic responses of the players in the market. For the asymmetric case, the existence of a sparse structure that captures bulk of the benefits of full flexibility under the partial postponement scenario and the supply disruptions scenario can also be challenging to prove. Moreover, our results on the fragility of systems exposed to supply disruptions can use some analytical strengthening. We leave these issues for future research.

# BIBLIOGRAPHY


[1] Chrysler Vice President inducted into Shingo Academy. *Assembly Magazine*, May 2005.

[2] The 100 top brands: Here's how Interbrand calculates the power in a name. *Business Week*, August 2007.

[3] O. Z. Aksin and F. Karaesmen. Characterizing the performance of process flexibility structures. *Operations Research Letters*, 35(4):477–484, 2007.

[4] W. Alderson. Marketing efficiency and the principle of postponement. *Cost and Profit Outlook*, September 1950.

[5] K. Anand and H. Mendelson. Postponement and information in a supply chain. Technical Report, Northwestern University, July 1998.

[6] A. Bassamboo, R. Randhawa, and J. A. Van Mieghem. A little flexibility is all you need: Asymptotic optimality of tailored chaining and pairing in queuing systems. Working paper, 2008.

[7] S. Benjafaar. Modeling and analysis of congestion in the design of facility layouts. *Management Science*, 48(5):679–704, 2002.

[8] D. Bertsimas and J. Tsitsiklis. *Introduction to Linear Optimization.* Athena Scientific, 1997.

[9] E. Bish, A. Muriel, and S. Biller. Managing flexible capacity in a make-to-order environment. *Management Science*, 51:167–180, 2005.

[10] E. Bish and Q. Wang. Optimal investment strategies for flexible resources, considering pricing and correlated demands. *Operations Research*, 52(954-964):6, 2004.

[11] N. Boudette. Chrysler gains edge by giving new flexibility to its factories. *The Wall Street Journal*, April 2006.

[12] K. Brack. Ripple effect from GM strike build. *Industrial Distribution*, 87(8):19, 1998.

[13] M. Brusco and T. Johns. Staffing a multiskilled workforce with varying levels of productivity: An analysis of cross-training policies. *Decision Sciences*, 29(2):499–515, 1998.

[14] G. Cachon and C. Terwiesch. *Matching Supply with Demand: An Introduction to Operations Management.* McGraw-Hill, 2006.

[15] S. Chopra, G. Reinhardt, and U. Mohan. The importance of decoupling recurrent and disruption risks in a supply chain. *Naval Research Logistics*, 54(5):544–555, 2007.

[16] M. Chou, G. Chua, C.P. Teo, and H. Zheng. Design for process flexibility: Efficiency of the long chain and sparse structure. *Operations Research*, To appear.

[17] M. Chou, C. P. Teo, and H. Zheng. Process flexibility revisited: Graph expander and its applications. Working Paper, 2008.

[18] D. P. de Farias and B. Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Math. Oper. Res.*, 29(3):462–478, 2004.

[19] K. DesMarteau. Leading the way in changing times. *Bobbin*, 41(2):48–54, 1999.

[20] G. Eppen. Effects on centralization on expected costs in a multilocation newsboy problem. *Management Science*, 25(5):498–501, 1979.

[21] B.Y. Feng. *100 Years of Li & Fung: Rise from Family Business to Multinational*. Thomson Learning, 2007.

[22] C. H. Fine and R. M. Freund. Optimal investment in product-flexible manufacturing capacity. *Management Science*, 36(4):449–466, 1990.

[23] M. Fisher and A. Raman. Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research*, 44(1):87–99, 1996.

[24] V. Fung, W. Fung, and Y. Wind. Competing in a flat world: The perils and promise of global supply chains. *ChangeThis*, 40:1–9, November 2007.

[25] V. Fung, W. Fung, and Y. Wind. *Competing in a Flat World: Building Enterprises for a Borderless World*. Wharton School Publishing, 2008.

[26] G. Gallego and R. Phillips. Revenue management of flexible products. *Manufacturing & Service Operations Management*, 6(4):321–337, 2004.

[27] S. C. Graves and B. T. Tomlin. Process flexibility in supply chain. *Management Science*, 49(7):907–919, 2003.

[28] S. Gurumurthi and S. Benjaafar. Modeling and analysis of flexible queueing systems. *Naval Research Logistics*, 51:755–782, 2004.

[29] J.H. Hammond and A. Raman. Sport Obermeyer, Ltd. *Harvard Business School Case*, 9-695-022:1–21, 1996. Boston, MA.

[30] W. J. Hopp, E. Tekin, and M. P. Van Oyen. Benefits of skill chaining in production lines with cross-trained workers. *Management Science*, 50(1):83–98, 2004.

[31] S. M. Iravani, M. P. Van Oyen, and K. T. Sims. Structural flexibility: A new perspective on the design of manufacturing and service operations. *Management Science*, 51(2):151–166, 2005.

[32] W. C. Jordan and S. C. Graves. Principles on the benefits of manufacturing process fexibility. *Management Science*, 41(4):577–594, 1995.

[33] R.M. Karp, R. Motwani, and N. Nisan. Probabilistic analysis of network flow algorithms. *Mathematics of Operations Research*, 18(1):71–97, 1993.

[34] M. Lahmar, H. Ergan, and S. Benjaafar. Resequencing and feature assignment on an automated assembly line. *IEEE Transactions on Robotics and Automation*, 19(1):89–102, 2003.

[35] T. LaSorda. Chrysler Group Update. January 2006.

[36] A. Latour. Trial by fire: A blaze in Albuquerque sets off major crisis for cell-phone giants – Nokia handles supply chain shock with aplomb as Ericsson of Sweden gets burned – Was Sisu the difference? *Wall Street Journal*, January 2001.

[37] H.L. Lee, C.A. Billington, and B. Carter. Hewlett Packard gains control of inventory and service through design for localization. *Interfaces*, 23(4):1–11, 1993.

[38] H.L. Lee and C.S. Tang. Modeling the costs and benefits of delayed product differentiation. *Management Science*, 43(1):40–53, 1997.

[39] R. Lien, S. M. Iravani, K. Smilowitz, and M. Tzur. Efficient and robust design for transshipment networks. Working Paper, 2005.

[40] M. Lim, A. Bassamboo, S. Chopra, and M. Daskin. Flexibility and fragility: Supply chain network design with disruption risks. Working Paper, 2008.

[41] R. Lyons, R. Pemantle, and Y. Peres. Resistance bounds for first-passage percolation and maximum flow. *Journal of Combinatorial Theory Ser. A*, 86(1):158–168.

[42] A.W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications*. Academic Press, 1979. Volume 143 in Mathematics in Science and Engineering Series.

[43] D. McCutcheon. Flexible manufacturing: IBM's bromont semiconductor packaging plant. *Canadian Electronics*, 19(7):26, 2004.

[44] J. Mouawad. Katrinas shock to the system. *New York Times*, September 2005.

[45] A. Muriel, A. Somasundaram, and Y. Zhang. Impact of partial manufacturing flexibility on production variability. *Manufacturing & Service Operations Management*, 8(2):192–205, 2006.

[46] S. Ross. *Introduction to Probability Models*. Academic Press, 2003.

[47] A. K. Sethi and S. P. Sethi. Flexibility in manufacturing: A survey. *The International Journal of Flexible Manufacturing Systems*, 2:289–328, 1990.

[48] S. Signorelli and J. L. Heskett. Benetton (a) and (b). *Harvard Business School Case*, 9-685-014:1–20, 1984. Boston, MA.

[49] N. Slack. The flexibility of manufacturing systems. *International Journal of Operations and Production Management*, 7(4):35–45, 1987.

[50] L. Snyder and Z.-J. Shen. Supply and demand uncertainty in multi-echelon supply chains. Working Paper, 2006.

[51] J. M. Swaminathan and H. L. Lee. Design for postponement. A. G. de Kok and S. C. Graves, eds. *Handbook of OR/MS in Supply Chain Management*. Chap. 5. Elsevier, Amsterdam, The Netherlands., 2003.

[52] B. Tomlin and Y. Wang. On the value of mix flexibility and dual sourcing in unreliable newsvendor networks. *Manufacturing & Service Operations Management*, 7(1):37–57, 2005.

[53] J. Van Biesebroeck. Flexible manufacturing in the North-American automobile industry. Working Paper, 2004.

[54] J. Van Biesebroeck. Complementarities in automobile production. *Journal of Applied Econometrics*, 22(7):1315–1345, December 2007.

[55] J. A. Van Mieghem. Investment strategies for flexible resources. *Management Science*, 44(8):1071–1078, 1998.

[56] J.A. Van Mieghem and M. Dada. Price versus product postponement: Capacity and competition. *Management Science*, 45:1632–1649, 1999.

[57] R. B. Wallace and W. Whitt. A staffing algorithm for call centers with skill-based routing. *Manufacturing and Service Operations Management*, 7(4):276–294, 2005.

[58] D. Z. Yu, S.i Y. Tang, H. Shen, and J. Niederhoff. On benefits of operational flexibility in a distribution network with transshipment. Working Paper, 2006.