Process Flexibility for Multi-Period Production Systems

Cong Shi
Industrial and Operations Engineering, University of Michigan, shicong@umich.edu
Yehua Wei
Carroll School of Management, Boston College, yehua.wei@bc.edu
Yuan Zhong
Booth School of Business, University of Chicago, Yuan.Zhong@chicagobooth.edu

We develop a theory for the design of process flexibility in a multi-period make-to-order production system. We propose and formalize a notion of "effective chaining" termed the Generalized Chaining Gap (GCG), which can be viewed as a natural extension of classical chaining structure from the process flexibility literature. Using the GCG, we prove that in a general system with high capacity utilization, one only needs a sparse flexibility structure with $m+n$ arcs to achieve similar performance as full flexibility, where m and n are equal to the number of plants and products in the system, respectively. The proof provides a simple and efficient algorithm for finding such sparse structures. Also, we show that the requirement of $m+n$ arcs is tight, by explicitly constructing systems in which even the best flexibility structure with $m+n-1$ arcs cannot achieve the same asymptotic performance as full flexibility. The goal of this paper is to make progress towards the better understanding of the key design principles of process flexibility structures in a multi-period environment.

Key words: process flexibility; flexible production; multi-period; capacity planning; chaining condition.
Received July 2017; revision received April, 2018; accepted July 2018 by Operations Research.

1. Introduction

In today's market, firms are often required to offer their customers more diverse product portfolios in order to keep ahead of the market competition. The increase in product offerings, however, would drive up the variabilities of product demands. To deal with this challenge, it has been observed that firms often adopt an operational strategy known as *process flexibility* (see Simchi-Levi (2010) and Cachon and Terwiesch (2011)). Also known as capacity pooling, process flexibility allows the firm to quickly change the production of different types of products from one plant to another with little penalty in time and cost, thereby adapting itself to reduce the operational costs under demand fluctuations.

Firms can choose to add different degrees of process flexibilities to their production system. For example, a production system has full flexibility if any of its plants is capable of producing any

of the product types (see Figure 1). Surprisingly, researchers have observed that the majority of the pooling benefits can be achieved with a small amount of flexibility. In the seminal paper of Jordan and Graves (1995), the authors observed that with the sparse *chaining* flexibility structure, one often obtains almost the same benefit as the fully flexible system. In a balanced system (same number of plants and product types), the chaining structure is also referred to as the long chain, as illustrated in Figure 1.

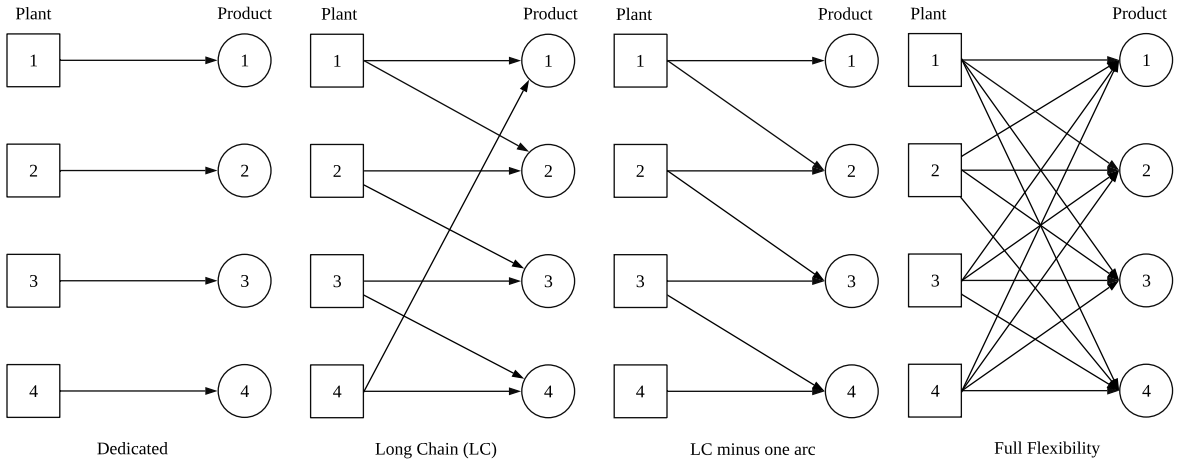

Figure 1 Examples of Process Flexibility Structures.

Much of the theoretical analysis on this topic focuses on a single-period model (see, e.g., Chou et al. (2010, 2011), Simchi-Levi and Wei (2012), Wang and Zhang (2015), Chen et al. (2015), Désir et al. (2016)). However, in practice, firms often operate in a multi-period setting, where the unsatisfied demand is backlogged into future time periods. This naturally leads to the following research question: in a multi-period make-to-order (MTO) environment, can a sparse flexibility structure achieve a performance that is close to that of the full flexibility structure? To answer this question, we first note that unlike the single-period setting, for any system with a non-trivial flexibility structure (structures that are neither dedicated nor fully flexible; see Figure 1 for illustrations), we are required to solve a complex multi-stage closed-loop dynamic optimization problem in the multi-period setting. Solving this problem optimally is computationally intractable, even for a small number of product types, because of the significant computational effort required over a large number of states in future time periods. To address this challenge, we apply a simple, well-known production policy known as the Maximum Weight (Max-Weight) policy (see, e.g., McKeown et al. (1999), Tassiulas and Ephremides (1992), Stolyar (2004), Dai and Lin (2005)) proposed in the

queueing literature, which has been shown to have near-optimal performance in broad classes of systems.

Because of the good performance properties of Max-Weight policy, for a flexibility structure, we can benchmark its performance under an optimal policy against that under Max-Weight policy, which often admits explicit bounds. This allows us to focus on the question of designing effective sparse flexibility structures. The first major contribution of our paper establishes that under realistic modeling assumptions, for any highly utilized system, i.e., a system whose utilization is close to 100%, there exists a sparse structure that performs almost as well as full flexibility. To prove this result, we introduce the concept of Generalized Chaining Gap (GCG), and use the theory of network flows to design a novel, efficient algorithm for constructing an effective sparse flexibility structure. The resulting structure uses at most $m+n$ arcs in a system with m plants and n products, and we show that it has the same asymptotic performance as the full flexibility structure, which consists of mn arcs.

Another major contribution of the paper is that we establish the necessity of having at least $m+n$ arcs for a flexibility structure to achieve a performance that is close to that of full flexibility. More specifically, we construct example systems in which any flexibility structure with at most $m+n-1$ arcs has a performance that is at least a constant factor away from the performance of full flexibility, however close the system utilization is to 100%. As a result, our analysis not only echoes "a little bit of flexibility goes a long way", a recurrent theme in the process flexibility literature, but also quantifies *how much* flexibility is needed in highly utilized multi-period make-to-order environment. An important step in characterizing the performance of these example systems uses a lower bound on the performance of any flexibility structure, which is derived using the state-of-the-art techniques from queueing theory. To the best of our knowledge, the lower bound we developed is new to both the process flexibility and the queueing literature.

Our modeling framework and results in this paper are closely related to so-called "parallel-server systems" (see, e.g., Stolyar (2004), Shah and Wischik (2012), Mandelbaum and Stolyar (2004), Harrison and López (1999), Gurvich and Whitt (2009)) that are widely studied in the queueing theory literature. Thus, we provide some remarks on the relationship and differences between our work and that literature. First, under a fixed flexibility structure, our model can be viewed as a discrete-time parallel-server system. However, rather than developing efficient control policies when the structure is given, which is a primary focus of the queueing literature, we focus on the separate problem of designing effective flexibility structures. Second, our design principle is closely related to, but differ in crucial ways from the so-called complete resource pooling (CRP) condition (see Stolyar (2004), Mandelbaum and Stolyar (2004), Harrison and López (1999), Gurvich and Whitt (2009), Ata and Kumar (2005)) from the literature on parallel-server systems. In the framework

developed in this paper, the CRP condition is equivalent to the existence of a positive GCG (that may, however, be very small), which can often be guaranteed by a tree flexibility structure. A flexibility structure with a positive GCG, however, need not achieve performance that is close to that of a fully flexible structure, as illustrated through our analysis and examples in §5. Indeed, for those examples, at least $m+n$ flexibility arcs are required to achieve similar performance as full flexibility, while there exist tree flexibility structures with exactly $m+n-1$ arcs satisfying CRP. This simple but important difference requires us to develop novel machinery for the constructions of effective flexibility structures, which provide valuable insights that cannot be inferred from the CRP condition alone.

The remainder of the paper is organized as follows. In the rest of this section, we provide a literature review and general notation. In §2, we describe the multi-period make-to-order system with process flexibility. In §3, we formalize the notion of the Generalized Chaining Gap (GCG), and use it to identify effective flexibility structures. In §4, we show that for general production systems with m plants and n products, it is possible to design effective flexibility structures with just $m+n$ production arcs. In §5, we show that in certain production systems, the $m+n$ production arcs is not only sufficient, but also necessary for designing effective flexibility structures. In §6, we perform numerical studies to investigate the robustness of our insights when systems deviate from our technical assumptions. In §7, we conclude our paper.

1.1. Literature Review

The study of process flexibility structures was first started by the seminal work by Jordan and Graves (1995). Recently, there has been much theoretical development to explain the power of chaining. In asymptotically large systems, Chou et al. (2010) developed a method to compute the average demand satisfied by the long chain. Chou et al. (2011) used graph expanders to show that there exists a sparse flexibility structure to achieve at least $(1-\varepsilon)$ performance of full flexibility for any $\varepsilon > 0$. Chen et al. (2015) used probabilistic graph expanders to strengthen the previous result (with high probability) using significantly fewer arcs, and Chen et al. (2016) then generalized the result for an asymmetrical and balanced system. Simchi-Levi and Wei (2012) identified a decomposition for the expected demand satisfied by the long chain and applied the decomposition to study its performance in finite systems. Wang and Zhang (2015) analyzed the long chain in a distributionally robust setting when only the first two moments of the demand are known. Désir et al. (2016) proved the optimality of the long chain among all connected structures that uses the same number of arcs. All these theoretical results developed so far were studied under the model proposed by Jordan and Graves (1995), which is effectively a single-period MTO system. We note that despite the recent developments, not much theory is known for the non-homogeneous,

finite sized single-period MTO systems. As a result, researchers have attempted to study non-homogeneous systems through either simulation (Deng and Shen (2013)), or different metrics and perspectives (Simchi-Levi and Wei (2015), Sheng et al. (2015)).

The work of Tanrisever et al. (2012) is one exception that studied chaining and partial flexibility structures under a multi-period MTO environment. They applied a sampling-based decomposition method to devise a feasible production scheduling policy, and used the policy to evaluate the effectiveness of different flexibility structures in simulations. In contrast, we theoretically demonstrate that certain sparse flexibilities are provably near-optimal under a much simpler production policy. The recent work of Asadpour et al. (2016) studied the allocation of flexible resource under the long chain structure. In their model, resources are depleted over time, which differs from our setting where resources have fixed capacities in each time period. Moreover, Asadpour et al. (2016) did not consider designing sparse flexibility structures in unbalanced systems (the number of product types is different from the number of resources).

Researchers have also studied the effectiveness of chaining and other partial structures in the context of queueing networks. Given the extensive literature in this area, we only review the most relevant works. In a series of works, Andradóttir et al. (2003, 2007, 2013) used the fluid model to study the capacity regions of flexible production systems; Hopp et al. (2004) studied worker skill-chaining in a U-shaped production line; Iravani et al. (2005) studied general partial flexibility structures in queueing networks; and finally, Tsitsiklis and Xu (2017) proved that queueing networks with expander properties simultaneously achieve large capacity region and vanishing queueing delay as the system size tends to infinity. With the exception of Tsitsiklis and Xu (2017), the aforementioned papers do not theoretically compare sparse flexibility with full flexibility. The key difference between Tsitsiklis and Xu (2017) and our work is that they studied large networks, while we focus on sparse flexibility structures in finite-size systems.

The Max-Weight policy used in this paper has been extensively studied in the queueing literature (see, e.g., McKeown et al. (1999), Tassiulas and Ephremides (1992), Stolyar (2004), Dai and Lin (2005)). Under the Complete Resource Pooling (CRP) condition, Max-Weight policy has been shown to achieve asymptotically optimal performance in the heavy-traffic limiting regime, using techniques such as weak convergence and diffusion approximation. Since the results in our paper concern performance in pre-limit systems, the approach that we take in analyzing the performance of GCG under Max-Weight policy follows Eryilmaz and Srikant (2012) closely, which is based on the simple but powerful idea of setting the drift of a Lyapunov function to zero in steady state, for non-limiting systems.

1.2. General Notation

Throughout this paper, symbols $\mathbb{R}$ and $\mathbb{R}_+$ are used to denote the set of reals and nonnegative reals, respectively. $\mathbb{Z}$ and $\mathbb{Z}_+$ are used to denote the set of integers and nonnegative integers, respectively. For a vector $\mathbf{x} = [x_i]$, $(\mathbf{x})^+$ is the vector whose components are given by $\max\{x_i, 0\}$. The vector and its scalar components are distinguished using bold letter and unbold letters respectively, e.g., given $\mathbf{x} \in \mathbb{R}^n$, the i-th entry of $\mathbf{x}$ is denoted using x_i. The vectors of all ones and all zeros are denoted by (boldface) $\mathbf{1}$ and $\mathbf{0}$, respectively. The inner product (a.k.a. scalar product) of two vectors is defined as $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^n x_i y_i$, $\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. The (Euclidean) norm of a vector in $\mathbb{R}^n$ is defined as $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$, $\forall\, \mathbf{x} \in \mathbb{R}^n$. For clarity, we often distinguish between a random variable and its realization using capital and lowercase letters, respectively. For two random variables X and Y, the notation $X \stackrel{d}{=} Y$ means that X and Y have the same probability distribution.

2. Multi-period Make-to-Order Systems with Process Flexibility

We describe a make-to-order (MTO) system with process flexibility over a planning horizon of T periods, where the time periods are indexed by $t = 1, \ldots, T$. The model naturally extends the single-period model considered in Jordan and Graves (1995) to a multi-period setting. We consider a stochastic make-to-order (MTO) system, where the manufacturer has $m \geq 1$ plants (each with capacity c_i, $i = 1, \ldots, m$) and $n \geq 1$ product types with some underlying process flexibility structure $\mathcal{A}$, which is represented by a set of arcs connecting the plant and the product nodes. The manufacturer can produce a product from a plant only if there is an arc connecting them. For each t, the demand vector in time period t is denoted by $\mathbf{D}(t) = [D_1(t), \ldots, D_n(t)]$, where $D_j(t)$ is the demand for product j in period t. We assume that $\mathbf{D}(t)$ are i.i.d. across time periods. Thus, we can use $\mathbf{D}$ to represent the *demand distribution* of the demand stream across time, i.e., $\mathbf{D}(t) \stackrel{d}{=} \mathbf{D}$ for each t. We also assume that $D_1(t), \ldots, D_n(t)$ are independent (though not necessarily identically distributed) across products.

For each $j \in \{1, 2, \ldots, n\}$, let λ_j be the expected demand in each period, i.e., $\mathbb{E}[D_j] = \lambda_j$, and let the *demand rate vector* be denoted as $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]$. Similarly, let $\sigma_j^2 = \mathrm{Var}[D_j]$ for each j, and we denote the variance vector by $\boldsymbol{\sigma}^2 = [\sigma_1^2, \ldots, \sigma_n^2]$. For notational convenience, we define $\Sigma^2(\boldsymbol{\sigma}^2) = \sum_{j=1}^n \sigma_j^2$, $\Lambda(\boldsymbol{\lambda}) = \sum_{j=1}^n \lambda_j$, and $C(\boldsymbol{c}) = \sum_{i=1}^m c_i$. When the context is clear, we often write Σ^2 for $\Sigma^2(\boldsymbol{\sigma}^2)$, Λ for $\Lambda(\boldsymbol{\lambda})$, and C for $C(\boldsymbol{c})$, to further simplify notation. We also let $c_{\min} = \min_{1 \leq i \leq m} c_i$, and $\lambda_{\min} = \min_{1 \leq j \leq n} \lambda_j$. Next, we state a set of regularity conditions that we assume throughout the paper.

ASSUMPTION 1. *There exist fixed positive constants l and u, such that for any capacity $\mathbf{c}$ and demand distribution $\mathbf{D}$, we have*

$$\text{Demand Conditions:} \quad P(D_j \le u) = 1,\ \lambda_j \ge l,\ \sigma_j \ge l,\ \forall 1 \le j \le n, \tag{1}$$

$$\text{Capacity Conditions:} \quad l \le c_i \le u,\ \forall 1 \le i \le m, \tag{2}$$

$$\text{Stability Condition:} \quad \Lambda(\boldsymbol{\lambda}) < C(\boldsymbol{c}). \tag{3}$$

We note that Assumption 1 is not restrictive to practical manufacturing settings. The demand for products are indeed bounded by their finite market size, while manufacturers do not produce products with a small amount of expected demand (Equation (1)). Also, manufacturing plants are typically required to have a certain level of capacities to operate efficiently (Equation (2)). Finally, the condition in Equation (3) requires that on average, there is more capacity than demand. As we shall see later, Equation (3) is a necessary and sufficient condition to guarantee that the system with full flexibility is *stable*, i.e., the system has finite long-run average backlogging cost.

Next, we introduce the concept of *average slack*, which is defined as

$$\zeta = \frac{C - \Lambda}{n}. \tag{4}$$

In this paper, we are especially interested in the regime where the total demand rate Λ is close to the total capacity C (equivalently, where ζ is small). This is often true in the context of flexible manufacturing. Indeed, it has been well documented (see, e.g., Cachon and Terwiesch (2011)) that flexibility is most valuable when capacity is approximately equal to the expected demand.

Finally, we use $\boldsymbol{\lambda}'$ to denote the *projection* of $\boldsymbol{\lambda}$ to the hyperplane defined by $\left\{\mathbf{g} \,|\, \sum_{j=1}^{n} g_j = C\right\}$. That is,

$$\boldsymbol{\lambda}' = \boldsymbol{\lambda} + \zeta \mathbf{e}, \quad \text{where } \mathbf{e} \text{ is the vector of 1's.} \tag{5}$$

2.1. Process Flexibility Structures and Production Polytope

For each $i \in \{1, 2, \ldots, m\}$, let $\mathcal{S}_i$ denote plant (node) i whose production capacity in each time period is c_i. Also, for each $j \in \{1, 2, \ldots, n\}$, let $\mathcal{T}_j$ denote product (node) j.

Process flexibility structure. We denote a *flexibility structure* by $\mathcal{A}$, which consists of a collection of arcs of the form $(\mathcal{S}_i, \mathcal{T}_j)$. Thus, $\mathcal{A}$ is the arc set of a bipartite graph with node partition $\{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$ and $\{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$. The production system is able to produce product j from plant i if and only if $(\mathcal{S}_i, \mathcal{T}_j) \in \mathcal{A}$.

Under a flexibility structure $\mathcal{A}$, let $N(\cdot)$ be the neighborhood function (note that here we suppress the dependence of $N(\cdot)$ on $\mathcal{A}$ to avoid overburdening the notation). The neighborhood function is defined as follows: for each $i = 1, 2, \ldots, m$, $N(\mathcal{S}_i) = \{\mathcal{T}_j \,|\, (\mathcal{S}_i, \mathcal{T}_j) \in \mathcal{A}\}$, and for each $j = 1, 2, \ldots, n$,

$N(\mathcal{T}_j) = \{\mathcal{S}_i \,|\, (\mathcal{S}_i, \mathcal{T}_j) \in \mathcal{A}\}$. Moreover, for any $\Omega \subseteq \{\mathcal{S}_1, \ldots, \mathcal{S}_m\} \cup \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$, $N(\Omega)$ is the set of all vertices that are neighbors to at least one node in Ω, i.e., $N(\Omega) = \cup_{\mathcal{X} \in \Omega} N(\mathcal{X})$. With the neighborhood function, we now formally define dedicated and full flexibility structure discussed in §1 (see Figure 1). A flexibility structure $\mathcal{A}$ is called a *dedicated* structure if no product can be produced from more than one plant, i.e., $|N(\mathcal{T}_j)| = 1$ for all $j = 1, \ldots, n$. A structure $\mathcal{A}$ is called a *full flexibility* structure if each product can be produced from all of the plants, i.e., $N(\mathcal{T}_j) = \{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$ for all $j = 1, \ldots, n$.

Production polytope. In each time period, the manager can decide how to allocate the flexible capacities $\boldsymbol{c}$ for production. We use $\mathbf{g} = [g_1, \ldots, g_n]$ to denote a generic production schedule vector, where g_j is the production amount for product j, for $j \in \{1, 2, \ldots, n\}$. The flexibility structure $\mathcal{A}$ places constraints on the production schedules, and we let $R(\mathcal{A})$ be the set of all feasible production schedules, and call it the *production polytope*. Because the production system does not change over time, the production polytope is time-invariant. It now follows that $R(\mathcal{A})$ is the set of all $\mathbf{g}$ such that there exists some vector $\mathbf{f} = [f_{i,j}] \in \mathbb{R}_+^{mn}$ where the following system of inequalities is satisfied.

$$\sum_{i=1}^{m} f_{i,j} = g_j, \forall j \in \{1, 2, \ldots, n\}, \tag{6}$$

$$\sum_{j=1}^{n} f_{i,j} \leq c_i, \forall i \in \{1, 2, \ldots, m\}, \tag{7}$$

$$f_{i,j} = 0, \forall (\mathcal{S}_i, \mathcal{T}_j) \notin \mathcal{A}. \tag{8}$$

For each i and j, $f_{i,j}$ (our decision variable) can be interpreted as the amount of production of product j at plant i. The first constraint (6) asserts that for each $j \in \{1, 2, \ldots, n\}$, the total production quantity g_j for product j is the sum of production quantities $f_{i,j}$ over all plants $i \in \{1, \ldots, m\}$. The second constraint (7) means that for each i, the total production quantity at plant i cannot exceed its capacity c_i. The last constraint (8) is subject to the underlying process flexibility structure $\mathcal{A}$.

2.2. Multi-Stage Optimization Model

Dynamics. Having defined the production polytope $R(\mathcal{A})$, we now describe the dynamics of the system under structure $\mathcal{A}$. In each time period, we assume that the production decision is made after observing the demand and the backlog vector at the beginning of the time period (or equivalently, at the end of the preceding time period). More specifically, let the backlog vector at the end of time period t be denoted by $\mathbf{B}(t) = [B_1(t), \ldots, B_n(t)]$, where $B_j(t)$ is the backlog for product j. Like product demands, the backlogs are stochastic, and a realized instance of the backlog vector at time t is denoted by $\mathbf{b}(t) = [b_1(t), \ldots, b_n(t)]$. For simplicity, we assume that $\mathbf{B}(0) = \mathbf{0}$ almost

surely, i.e., the system is initially empty. In each time period $t \in \{1, \ldots, T\}$, the sequence of events is described as follows.

(a) The manager observes the starting backlog levels $\mathbf{b}(t-1) = [b_1(t-1), \ldots, b_n(t-1)]$ before the demand occurs in period t. Then, the demand vector $\mathbf{D}(t)$ realizes to be $\mathbf{d}(t)$ in our MTO system. The manager observes $\mathbf{d}(t)$ and updates the backlog levels as

$$\mathbf{b}'(t) = \mathbf{b}(t-1) + \mathbf{d}(t). \tag{9}$$

In this paper, we refer to $\mathbf{b}'(t)$ as the *in-period backlog* during period t.

(b) The manager then decides to produce $\mathbf{g}(t) = [g_1(t), \ldots, g_n(t)] \in R(\mathcal{A})$, which satisfies the production constraints (6)–(8), and the backlog levels after production become $\mathbf{b}(t) = \big(\mathbf{b}'(t) - \mathbf{g}(t)\big)^+$. In this paper, due to the assumed across-time independence of demands, we restrict our attention to closed-loop feasible policy π, which is determined by a sequence of (measurable) functions $\mathbf{g}(t) = \pi_t(\mathbf{b}'(t))$, $t = 1, \ldots, T$, mapping in-period backlog $\mathbf{b}'(t)$ (state) into production schedule $\mathbf{g}(t) \in R(\mathcal{A})$ (see Bertsekas and Shreve (2007)). The state transition is written as

$$\mathbf{b}(t) = \big(\mathbf{b}(t-1) + \mathbf{d}(t) - \mathbf{g}(t)\big)^+. \tag{10}$$

Performance measure. We assume uniform (per-unit) backlogging cost across different products, and, without loss of generality, we set it to be 1. We remark that uniform cost or profit is a common assumption used in process flexibility literature; see e.g., Jordan and Graves (1995), Chou et al. (2010), etc. To study the performance of a flexible system under a single period model, Jordan and Graves (1995) analyzed the quantity

$$\min_{\pi} \mathbb{E}\left[\sum_{j=1}^{n} B_j^{\pi}(1)\right], \tag{11}$$

where the optimal policy π can be solved through a max-flow problem on a bipartite network, specified by the flexibility structure $\mathcal{A}$. In contrast, we focus on multi-period models, and we are primarily interested in studying the minimum of the expected long-run average backlogging costs, i.e.,

$$\min_{\pi} \Gamma(\pi), \quad \text{where} \quad \Gamma(\pi) = \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{n} B_j^{\pi}(t)\right], \tag{12}$$

and π is a closed-loop feasible policy for the multi-period stochastic optimization model. Since we will often be comparing the optimal long-run average backlogging costs $\min_{\pi} \Gamma(\pi)$ under different structures $\mathcal{A}$, we write $BL(\mathcal{A}) = \min_{\pi} \Gamma(\pi)$ to denote the performance measure of flexibility structure $\mathcal{A}$. When $BL(\mathcal{A})$ is finite, the system with flexibility structure $\mathcal{A}$ is called *stable*. It is well

known in the queueing literature that the system with flexibility structure $\mathcal{A}$ is stable if and only if

$$\sum_{\mathcal{S}_i \in N(\Omega)} c_i > \sum_{\mathcal{T}_j \in \Omega} \lambda_j, \text{for all } \Omega \subseteq \{\mathcal{T}_1, \ldots, \mathcal{T}_n\},\ \Omega \neq \emptyset. \tag{13}$$

Details of the stability condition are provided §EC.1. Note that the system with full flexibility is stable if and only if $\Lambda < C$, which is precisely Equation (3) stated in Assumption 1.

When the stability condition for $\mathcal{A}$ is satisfied, the multi-period stochastic optimization model can be formulated as an infinite horizon dynamic programming (DP), with a state space consists of n-dimensional vectors. Unfortunately, even with moderate sizes of n, solving this DP optimally is computationally intractable, as the state space grows exponentially fast with n. This is well-known to be the *curse of dimensionality* (see e.g., Powell (2007)). Therefore, instead of solving the DP optimally, we leverage simple policies proposed in the queueing literature to study the stochastic optimization model.

Relationship with the parallel server system. We end this section by providing some remarks on the relationship between our model and the parallel server system model (see, e.g., Mandelbaum and Stolyar (2004), Stolyar (2004)). In the terminology of this paper, a discrete-time parallel server system has m servers (plants) and n queues (products). In each time period, a server is only allowed to serve one queue, and the number of type-j jobs that can be processed by server i is $\mu_{i,j}$. $\mu_{i,j}$ can be interpreted as service rates/capacities, and they are called *server-dependent* if for each i, there exists c_i such that $c_i \equiv \mu_{i,j}$ for all j. Under any given flexibility structure $\mathcal{A}$, our system can be viewed as a discrete-time parallel server system with server-dependent capacities, with the difference that we allow plant capacities to be shared among the products in any arbitrary manner. Another related model is the one considered in Gurvich and Whitt (2009), which is a "many-server" service system with multiple customer classes and *server pools*, where the service rates are pool dependent. In their model, demand rates and numbers of servers in each pool scale to infinity, whereas we consider finite-size systems.

3. Effective Flexibility Structures

3.1. Generalized Chaining Gap

We introduce the *Generalized Chaining Gap*, an important measure that we use to understand the effectiveness of flexibility structures. Recall the average slack $\zeta = (C - \Lambda)/n$ as defined in (4), and $\boldsymbol{\lambda}' = \boldsymbol{\lambda} + \zeta \boldsymbol{e}$ as defined in (5), which is the projection of $\boldsymbol{\lambda}$ to the plane defined by $\left\{\mathbf{g} \mid \sum_{j=1}^{n} g_j = C\right\}$. We note that $\sum_{i=1}^{m} c_i = \sum_{j=1}^{n} \lambda'_j = \sum_{j=1}^{n} (\lambda_j + \zeta)$.

DEFINITION 1. Fix a flexibility structure $\mathcal{A}$. Its *Generalized Chaining Gap* (GCG) is defined as

$$\eta \triangleq \min_{\Omega \subsetneq \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n\}, \Omega \neq \emptyset} \left\{ \sum_{\mathcal{S}_i \in N(\Omega)} c_i - \sum_{\mathcal{T}_j \in \Omega} \lambda'_j \right\}. \tag{14}$$

Here we provide some intuition behind GCG. First, note that if we increase the demand rate $\mathbb{E}[\mathbf{D}]$ to $\boldsymbol{\lambda}'$, then the total capacity utilization rate becomes 100% and the system (even with full flexibility) becomes unstable, i.e., the long-run average backlogging cost becomes infinity. Now, consider a fixed flexibility structure $\mathcal{A}$, and suppose that its GCG is strictly positive, i.e., $\eta > 0$. This implies that for any strict subset $\Omega \subsetneq \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n\}$, if we increase the demand rate $\mathbb{E}[\mathbf{D}]$ to $\boldsymbol{\lambda}'$, the total demand rate for products in Ω is less than the total capacity that can be used to produce products in Ω, which intuitively implies that the system can be "locally stable" (having finite backlogs) at Ω. Therefore, GCG (when it is positive) can be thought as measuring the strength of $\mathcal{A}$'s "local stability" for all product (strict) subsets, when demand rate is increased to the point where system itself becomes unstable. In the next subsection, we will show that the long-run average backlogging cost of a flexibility structure $\mathcal{A}$ can be formally analyzed using the notion of GCG.

Let us end this subsection by providing two remarks about GCG.

REMARK 1. For a flexibility structure $\mathcal{A}$, it is not difficult to show that the condition $\eta > 0$ is equivalent to the following:

CONDITION 1 *There exists a vector* $\mathbf{f} = [f_{i,j}] \in \mathbb{R}_+^{mn}$ *that satisfies* $\sum_j f_{i,j} = c_i$ *for all* i, $\sum_i f_{i,j} = \lambda'_j$ *for all* j, *such that the graph* $\mathcal{G} = \{(\mathcal{S}_i, \mathcal{T}_j) : f_{i,j} > 0, 1 \le i \le m, 1 \le j \le n\}$ *is connected and* $\mathcal{G} \subseteq \mathcal{A}$.

Since $\mathcal{G} \subseteq \mathcal{A}$, an immediate consequence is that if $\mathcal{A}$ is disconnected, then its GCG cannot be positive. We also note that Condition 1 is essentially the same as Assumption 2.4 of Gurvich and Whitt (2009), often referred to as the Complete Resource Pooling (CRP) condition in the queueing literature.

The next remark considers the GCG of the classical long chain structure in a balanced system (that has an equal number of plants and products; i.e., $m = n$) with uniform plant capacity and uniform demand rates. Formally, a structure $\mathcal{A}$ in an n-plant n-product system is a *long chain* if $\mathcal{A} = \{(\mathcal{S}_1, \mathcal{T}_1), (\mathcal{S}_1, \mathcal{T}_2), (\mathcal{S}_2, \mathcal{T}_2), (\mathcal{S}_2, \mathcal{T}_3), \ldots (\mathcal{S}_n, \mathcal{T}_n), (\mathcal{S}_n, \mathcal{T}_1)\}$ (see Figure 1 for an example).

REMARK 2. Let λ and c be two given positive constants with $\lambda < c$. Consider a balanced system of size n (i.e., n plants and n products) with $c_i = c$ and $\lambda_j = \lambda$ for all i and j. Then the GCG of the long chain is exactly c.

Remark 2 follows by noting that $\lambda'_j = c$ for $1 \le j \le n$. Thus, for any non-empty strict subset Ω of $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n\}$, $\sum_{\mathcal{S}_i \in N(\Omega)} c = c|N(\Omega)| = (|\Omega|+1)c$, while $\sum_{\mathcal{T}_j \in \Omega} \lambda'_j = |\Omega|c$.

We note that the notion of chaining was first introduced by Jordan and Graves (1995), under the description "a group of products and plants which are all connected, directly or indirectly, by product assignment decisions." While Jordan and Graves (1995) presented the long chain as an example of chaining in the balanced ($m = n$) system, it does not provide a formal definition of chaining in unbalanced systems. Therefore, the definition of Generalized Chaining Gap can be thought of as an extension of the chaining idea from Jordan and Graves (1995).

3.2. Bounding the Performance under GCG

Here we analyze the long-run average backlogging cost for flexibility structures with positive GCG. For a structure $\mathcal{A}$, we are interested in the performance measure $BL(\mathcal{A})$, where we recall that $BL(\mathcal{A}) = \min_\pi \Gamma(\pi)$, with $\Gamma(\pi)$ defined in (12). To do so, we leverage known results and techniques from queueing theory and upper bound $BL(\mathcal{A})$ by upper bounding $\Gamma(MW)$ under the well-known Max-Weight (MW) policy (McKeown et al. (1999), Tassiulas and Ephremides (1992), Stolyar (2004), Dai and Lin (2005), etc). Since our model (under a given structure $\mathcal{A}$) is slightly different from traditional discrete-time parallel server systems, we provide a full description of the Max-Weight policy for completeness.

Definition 2. Under the Max-Weight policy, at period t, given that the last period backlog is $\mathbf{b}(t-1)$ and current period demand is $\mathbf{d}(t)$, the policy determines the production schedule $\mathbf{g}$ by solving following optimization problem:

$$\begin{aligned} & \max \sum_{j=1}^{n} (b_j(t-1) + d_j(t)) \cdot g_j && \text{(Opt-MW)} \\ \text{s.t.} \quad & g_j \le b_j(t-1) + d_j(t), \\ & \mathbf{g} \in R(\mathcal{A}). \end{aligned}$$

We note that Problem Opt-MW may have multiple optimal solutions. For the sake of simplicity, we assume that the Max-Weight policy applies some arbitrary fixed tie-breaking rule when such cases arise.

We now provide some results to analyze the performance of flexibility structures with positive GCG. These results can be derived using techniques from Eryilmaz and Srikant (2012), whose proofs are provided in EC.2 for completeness. Recall that Σ^2 is defined as the sum of variance for products in the system.

PROPOSITION 1. *Let $\Lambda < C$, and let $\mathcal{A}$ be a flexibility structure with $\eta > 0$. Then,*

$$BL(\mathcal{A}) \le \Gamma(MW) \le \frac{\Sigma^2}{2n\zeta} + \frac{K_1 + \eta K_2}{\eta\sqrt{\zeta}}, \tag{15}$$

where $\Gamma(MW)$ is the long-run average total backlogging cost under the Max-Weight policy, and $K_1 = K_1(l,u)$ and $K_2 = K_2(l,u)$ are positive constants that only depend on l and u in a continuous manner.

PROPOSITION 2. *Let $\Lambda < C$ and consider the fully flexible structure. Then, we have*

$$\frac{\Sigma^2}{2n\zeta} - \frac{C - n\zeta}{2} \le BL(\mathcal{F}) \le \frac{\Sigma^2}{2n\zeta} + \frac{C - \Lambda}{2}, \tag{16}$$

where $BL(\mathcal{F})$ denotes the (optimal) performance of the fully flexible structure.

COROLLARY 1. *The performance of any flexibility structure $\mathcal{A}$ can be lower bounded as*

$$BL(\mathcal{A}) \ge \frac{\Sigma^2}{2n\zeta} - \frac{C - n\zeta}{2}. \tag{17}$$

An immediate consequence of Proposition 1 and Corollary 1 is that when capacity utilization is high, i.e., $\zeta \approx 0$, the ratio between the performance of the long chain to that of full flexibility approaches to one.

COROLLARY 2. *Consider a balanced system of fixed size n with $c_i = c_j$, $\lambda_i = \lambda_j$ for all $1 \le i, j \le n$. Under Assumption 1, there exists a constant $K = K(l,u) > 0$ that depends only on l and u continuously, such that for all sufficiently small ζ,*

$$\frac{BL(\mathcal{LC})}{BL(\mathcal{F})} \le 1 + K\sqrt{\zeta}, \tag{18}$$

where $BL(\mathcal{LC})$ and $BL(\mathcal{F})$ denote the long-run average backlogging costs (under the optimal policy) of long chain and full flexibility, respectively.

The proof of Corollary 2 can be found in §EC.2. It has been well documented that in a single period system, the long chain performs almost as well as full flexibility (Jordan and Graves 1995, Chou et al. 2010, Simchi-Levi and Wei 2012). Thus, one can view Corollary 2 as an analogous result for the long chain in a multi-period environment. Unlike the single period system literature, where the long chain achieves a close, but strictly inferior performance when compared to full flexibility, Corollary 2 illustrates that the long chain in our multi-period environment is *asymptotically* close to full flexibility as the capacity slack ζ approaches to zero.

Corollary 2 naturally leads to the question of whether there exists a similar asymptotic result for general systems that are not balanced and symmetric. This question is investigated in full detail in the next section.

4. Designing Sparse Flexibile Structures in Unbalanced Systems

In this section, we consider the question of designing sparse structures under the unbalanced, asymmetric systems. We show that in an unbalanced system with m plants, and n products, it is possible to create an effective flexibility structure with $m+n$ production arcs, compared to mn arcs required by the full flexibility structure. More specifically, similar to the performance of long chains in balanced systems (Corollary 2), we show that with $m+n$ production arcs, one can construct a flexibility structure that performs asymptotically close to full flexibility, when the capacity slack is small.

THEOREM 1. *Under Assumption 1, there exists $K = K(l,u) > 0$ and a process flexibility structure $\mathcal{A}$ with $m+n$ production arcs ($|\mathcal{A}| = m+n$), such that for all sufficiently small $\zeta > 0$,*

$$\frac{BL(\mathcal{A})}{BL(\mathcal{F})} \leq 1 + K\sqrt{\zeta}, \tag{19}$$

where $BL(\mathcal{A})$ and $BL(\mathcal{F})$ denote the long-run average backlogging costs (under the optimal policy) of $\mathcal{A}$ and full flexibility, respectively.

Let us first outline the ideas for proving Theorem 1. Like the balanced symmetric system (Corollary 2), the proof of Theorem 1 also applies the performance bounds on the flexibility structures with positive GCG derived in Proposition 1 and Corollary 1. However, there is not a clear notion of the long chain in asymmetric unbalanced systems, and one needs to carefully consider how to design a flexibility structure $\mathcal{A}$ with $m+n$ arcs.

It may be tempting to conjecture that any structure $\mathcal{A}$ with strictly positive GCG is "sufficient", as it is the condition required for applying Proposition 1. Interestingly, strictly positive GCG is not enough to satisfy Equation (19). In fact, in §5, we prove that there exists a flexibility structure $\mathcal{A}$ with strictly positive GCG that does not achieve asymptotic optimality, in the sense of $BL(\mathcal{A})/BL(\mathcal{F}) \to 1$ as $\zeta \to 0$. This is because under this structure, for a sequence of systems whose average slack ζ approaches 0, even though the GCG η remains positive, it approaches 0 at the same rate as ζ. As a result, to prove Theorem 1, we need a stronger condition that ensures the GCG of our sparse flexibility structure $\mathcal{A}$ is not only positive, but also sufficiently large. The following subsection describes a procedure that generates "sufficient" flexibility structures with just $m+n$ arcs.

4.1. Constructing Sufficient GCG

PROPOSITION 3. *Consider a system with m plants, n products, capacity vector $\mathbf{c}$ and demand rate vector $\boldsymbol{\lambda}$. There exists an algorithm that generates flexibility structure $\mathcal{A}$ with $m+n$ production arcs ($|\mathcal{A}| = m+n$), such that its GCG is at least δ, where*

$$\delta \triangleq \frac{\min\{\lambda'_{\min}, c_{\min}\}}{\min\{m,n\}}. \tag{20}$$

Moreover, the algorithm terminates in $O(m+n)$ operations.

To prove Proposition 3, recall that for $\mathcal{A}$ to have a GCG of at least δ, we must have that for any nonempty subset $\Omega \subsetneq \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n\}$,

$$\sum_{\mathcal{S}_i \in N(\Omega)} c_i - \sum_{\mathcal{T}_j \in \Omega} \lambda'_j \geq \delta. \tag{21}$$

To find the desired flexibility structure, we propose a two-step *partition and join* procedure. In the first step, we find an acyclic flexibility structure $\mathcal{A}'$ that *partitions* the plant and product nodes into k disjoint components, such that in each component of $\mathcal{A}'$, Equation (21) is *almost* satisfied (cf. Lemma 2). In the second step, we add arcs to *join* all of the components in $\mathcal{A}'$ together, and complete the flexibility structure $\mathcal{A}$ with a GCG of at least δ.

The *partition* step of our procedure is presented as Algorithm 1, which returns flexibility structure $\mathcal{A}'$. Algorithm 1 also returns a flow $\mathbf{f}$ on $\mathcal{A}'$, which will also be used for our analysis.

Algorithm 1 Finding flexibility structure $\mathcal{A}'$

Input: $\mathbf{c}$, $\boldsymbol{\lambda}'$ and δ.

Set the initial values to $\mathcal{A}' = \emptyset$, $i = j = 1$, $s = c_1$ and $t = \lambda'_1$.

while $i < m$ or $j < n$ **do**

  Set $f_{i,j} = \min\{s,t\}$, and $\mathcal{A}' = \mathcal{A}' \cup \{\mathcal{S}_i, \mathcal{T}_j\}$.

  if $|s-t| < \delta$ **then** set $i = i+1$, $j = j+1$, $s = c_i$ and $t = \lambda'_j$.

  else if $s - t \geq \delta$ **then** set $j = j+1$, $s = s-t$, $t = \lambda'_j$.

  else if $t - s \geq \delta$ **then** set $i = i+1$, $t = t-s$, $s = c_i$.

  end if

end while

Return $\mathcal{A}'$ and $\mathbf{f}$.

To provide some intuition for Algorithm 1, it is useful to consider the case where δ takes value 0 in the algorithm. In this case, Algorithm 1 solves a static max-flow problem with capacity $\mathbf{c}$ in

the plant nodes and demand $\boldsymbol{\lambda}'$ in the product nodes, by adding appropriate arcs to form $\mathcal{A}'$ and to greedily exhaust the capacities of plant nodes from 1 to m to satisfy the demand λ'_j of product nodes. When $\delta > 0$, Algorithm 1 is a similar greedy algorithm that ensures each arc in $\mathcal{A}'$ has a flow of size at least δ.

We now describe some useful properties of the flexibility structure $\mathcal{A}'$ returned from Algorithm 1, when $\delta = \frac{\min\{\lambda'_{\min}, c_{\min}\}}{\min\{m,n\}}$. Suppose that $\mathcal{A}'$ has k connected components, $\mathscr{C}_1, \ldots, \mathscr{C}_k$. Then, by definition of Algorithm 1, we have that (i) each arc in $\mathcal{A}'$ has a flow of size at least δ, (ii) each component $\mathscr{C}_l$ is a tree, and (iii) each component contains at least 1 plant node and 1 product node. The first two properties are immediate; and the third property is proved in the next lemma.

LEMMA 1. *For each $l \in \{1, 2, \ldots, k\}$, component $\mathscr{C}_l$ of $\mathcal{A}'$ contains at least 1 plant node and 1 product node.*

Proof of Lemma 1. We begin the proof by providing the following observation about the flexibility structure $\mathcal{A}'$. First, by construction, for all components that do not contain at least 1 plant node and 1 product node, they must all be isolated nodes, which are either all plant nodes or all product nodes.

We now prove Lemma 1 by contradiction. For each $l \in \{1, 2, \ldots, k\}$, let Δ_l be the difference between the aggregate capacity and the aggregate demand (defined by $\boldsymbol{\lambda}'$) in $\mathscr{C}_l$. Then, it is easy to see that $\sum_{l=1}^{k} \Delta_l = 0$, since $\sum_{i=1}^{m} c_i = \sum_{j=1}^{n} \lambda'_j$. Without loss of generality, suppose that for $l \in \{1, 2, \ldots, k'\}$, $\mathscr{C}_l$ contains at least 1 plant node and 1 product node, and for all $l' \in \{k'+1, \ldots, k\}$, $\mathscr{C}_{l'}$ is a singleton. Then, $k' \leq \min\{m, n\}$, and by way of contradiction, $k' < k$.

For $l \in \{1, 2, \ldots, k'\}$, $|\Delta_l| < \delta$, so

$$|\Delta_1 + \cdots + \Delta_{k'}| \leq |\Delta_1| + \cdots + |\Delta_{k'}| < k'\delta \leq \min\{m, n\}\delta = \min\{\lambda'_{\min}, c_{\min}\}.$$

For $l' > k'$, since $\mathscr{C}_{l'}$ is a singleton, $|\Delta_{l'}| \geq \min\{\lambda'_{\min}, c_{\min}\}$. Furthermore, since $\mathscr{C}_{l'}$ are either all plant nodes or product nodes,

$$|\Delta_{k'+1} + \cdots + \Delta_k| = |\Delta_{k'+1}| + \cdots + |\Delta_k| \geq \min\{\lambda'_{\min}, c_{\min}\}.$$

But this is a contradiction, since

$$|\Delta_{k'+1} + \cdots + \Delta_k| = |\Delta_1 + \cdots + \Delta_k - (\Delta_1 + \cdots + \Delta_{k'})| = |0 - (\Delta_1 + \cdots + \Delta_{k'})| < \min\{\lambda'_{\min}, c_{\min}\}.$$

This shows that each component of $\mathcal{A}'$ contains at least 1 plant node and 1 product node. □

By Lemma 1, we can relabel the components and the nodes so that there are integers $1 \le i_1 < i_2 < \ldots < i_k = m$, and $1 \le j_1 < j_2 < \ldots < j_k = n$ such that

$$\{\mathcal{T}_1 \ldots \mathcal{T}_{j_1}\} \in \mathscr{C}_1, \{\mathcal{T}_{j_1+1} \ldots \mathcal{T}_{j_2}\} \in \mathscr{C}_2, \ldots, \{\mathcal{T}_{j_{k-1}+1} \ldots \mathcal{T}_n\} \in \mathscr{C}_k,$$
$$\text{and } \{\mathcal{S}_1 \ldots \mathcal{S}_{i_1}\} \in \mathscr{C}_1, \{\mathcal{S}_{i_1+1} \ldots \mathcal{S}_{i_2}\} \in \mathscr{C}_2, \ldots, \{\mathcal{S}_{i_{k-1}+1} \ldots \mathcal{S}_m\} \in \mathscr{C}_k.$$

In addition, for each $l \in \{1, 2, \ldots, k\}$, we now formally define Δ_l, the difference between the aggregate capacity and aggregate demand (defined by $\boldsymbol{\lambda}'$) in $\mathscr{C}_l$ as $\Delta_l = \sum_{i=i_{l-1}+1}^{i_l} c_i - \sum_{j=j_{l-1}+1}^{j_l} \lambda'_j$. By the definition of Algorithm 1, we must have that $|\Delta_l| < \delta$ for all $l \in \{1, 2, \ldots, k-1\}$, and $\sum_{l=1}^{k} \Delta_l = 0$. This implies that for any $L \subsetneq \{1, \ldots, k\}$, we have

$$\sum_{l \in L} \Delta_l \le \min \left\{ \Big| \sum_{l \in L} \Delta_l \Big|, \Big| \sum_{l \notin L} \Delta_l \Big| \right\} \le (k-1)\delta. \tag{22}$$

The next lemma shows that (21) is satisfied for almost all product subsets in $\mathscr{C}_l$ for any $l = 1, \ldots k$.

LEMMA 2. *For each $l \in \{1, 2, \ldots, k\}$, and any $\Omega \subseteq \{\mathcal{T}_{j_{l-1}+1} \ldots \mathcal{T}_{j_l}\}$, let $N'(\Omega)$ denote the set of neighbors of Ω under $\mathcal{A}'$. Then we have*

$$\sum_{\mathcal{S}_i \in N'(\Omega)} c_i - \sum_{\mathcal{T}_j \in \Omega} \lambda'_j = 0 \qquad \text{if } \Omega = \emptyset; \tag{23}$$
$$\sum_{\mathcal{S}_i \in N'(\Omega)} c_i - \sum_{\mathcal{T}_j \in \Omega} \lambda'_j = \Delta_l \qquad \text{if } \Omega = \{\mathcal{T}_{j_{l-1}+1} \ldots \mathcal{T}_{j_l}\}; \tag{24}$$
$$\sum_{\mathcal{S}_i \in N'(\Omega)} c_i - \sum_{\mathcal{T}_j \in \Omega} \lambda'_j \ge \Delta_l + \delta \qquad \text{if } \emptyset \ne \Omega \subsetneq \{\mathcal{T}_{j_{l-1}+1} \ldots \mathcal{T}_{j_l}\}, \mathcal{T}_{j_l} \in \Omega; \tag{25}$$
$$\sum_{\mathcal{S}_i \in N'(\Omega)} c_i - \sum_{\mathcal{T}_j \in \Omega} \lambda'_j \ge \delta \qquad \text{if } \emptyset \ne \Omega \subsetneq \{\mathcal{T}_{j_{l-1}+1} \ldots \mathcal{T}_{j_l}\}, \mathcal{T}_{j_l} \notin \Omega. \tag{26}$$

Proof of Lemma 2. It is clear that (23) and (24) follow directly by definition. To prove (25), consider the flow problem on $\mathscr{C}_l$, where plant node $\mathcal{S}_i \in \mathscr{C}_l$ has capacity c_i and product node $\mathcal{T}_j \in \mathscr{C}_l$ has demand λ'_j for all $j \ne j_l$, and product node $\mathcal{T}_{j_l}$ has demand $\lambda'_{j_l} + \Delta_l$. Let $\mathbf{f}^*$ be a flow on $\mathscr{C}_l$ such that

$$f^*_{i,j} = f_{i,j}, \forall (\mathcal{S}_i, \mathcal{T}_j) \in \mathscr{C}_l, (\mathcal{S}_i, \mathcal{T}_j) \ne (\mathcal{S}_{i_l}, \mathcal{T}_{j_l}),$$
$$f^*_{i_l,j_l} = f_{i_l,j_l} + \Delta_l, \text{ if } \Delta_l > 0, \text{ and } f^*_{i_l,j_l} = f_{i_l,j_l}, \text{ if } \Delta_l < 0,$$

where we recall that $\mathbf{f} = (f_{i,j})$ is the flow returned from Algorithm 1. Then, by the definition of Algorithm 1, $\mathbf{f}^*$ is a flow that satisfies all of the demand for the flow problem on $\mathscr{C}_l$ under consideration. Note that for any $\Omega \subsetneq \{\mathcal{T}_{j_{l-1}+1} \ldots \mathcal{T}_{j_l}\}$, there must exist a plant node that sends at

least δ units of flow from a plant node $\mathcal{S}_i$ in $N'(\Omega)$ to another product node $\mathcal{T}_j$ not in Ω. Therefore, we have

$$\sum_{\mathcal{S}_i \in N'(\Omega)} c_i \geq \sum_{\mathcal{S}_i \in N'(\Omega)} \sum_{\mathcal{T}_j \in \Omega} f^*_{i,j} \geq \sum_{\mathcal{T}_j \in \Omega} \lambda'_j + \Delta_l + \delta.$$

Finally, to prove (26), note that flow $\mathbf{f}$ sends λ'_j units of flow to product node $\mathcal{T}_j \in \mathscr{C}_l$ for any $j \neq j_l$. Moreover, if $\Omega \subsetneq \mathscr{C}_l$, there must exist a plant node $\mathcal{S}_i$ in $N'(\Omega)$ that sends at least δ units of flow to another product node $\mathcal{T}_j$ not in Ω. Therefore, we have

$$\sum_{\mathcal{S}_i \in N'(\Omega)} c_i \geq \sum_{\mathcal{S}_i \in N'(\Omega)} \sum_{\mathcal{T}_j \in \Omega} f_{i,j} \geq \sum_{\mathcal{T}_j \in \Omega} \lambda'_j + \delta,$$

which proves (26). □

Lemma 2 shows (21) is satisfied for all nonempty product subsets in $\mathscr{C}_l$ not containing $\mathcal{T}_{j_l}$. In the next lemma (Lemma 3), we describe the *join* step of our procedure, which adds k arcs to $\mathcal{A}'$ to create structure $\mathcal{A}$, which connects components $\mathscr{C}_1, \ldots, \mathscr{C}_k$ and provides us with the desired sparse structure.

LEMMA 3. *Consider the flexibility structure $\mathcal{A}$ defined as*

$$\mathcal{A} = \mathcal{A}' \cup \{(\mathcal{S}_{i_2}, \mathcal{T}_{j_1}), (\mathcal{S}_{i_3}, \mathcal{T}_{j_2}), \ldots, (\mathcal{S}_{i_k}, \mathcal{T}_{j_{k-1}}), (\mathcal{S}_{i_1}, \mathcal{T}_{j_k})\}. \tag{27}$$

Then, for any $\emptyset \neq \Omega \subsetneq \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$, we have

$$\sum_{\mathcal{S}_i \in N(\Omega)} c_i - \sum_{\mathcal{T}_j \in \Omega} \lambda'_j \geq \delta,$$

where $N(\Omega)$ denotes the set of neighbors of Ω under $\mathcal{A}$, and $\delta = \frac{\min\{\lambda'_{\min}, c_{\min}\}}{\min\{m,n\}}$.

Proof of Lemma 3. Let Ω be a nonempty proper subset of $\{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$. Similar to Lemma 2, let $N'(\Omega)$ denote the set of neighbors of Ω under $\mathcal{A}'$. Note that because each component $\mathscr{C}_l$ in $\mathcal{A}'$ has at least 1 plant node and 1 product node, we have that $k \leq \min\{m, n\}$, which implies that

$$\delta \leq \frac{\min\{\lambda'_{\min}, c_{\min}\}}{k}. \tag{28}$$

Without loss of generality, we may assume that the subgraph of $\mathcal{A}$ induced by Ω is connected. Also, for notational convenience, we assume that $i_{k+1} = i_1$. To prove Lemma 3, we will look at four different cases that cover all of the possibilities. The four cases are: (i) there exists some $l' \in \{1, 2, \ldots, k\}$ such that $\mathcal{T}_{j_{l'}} \in \Omega$ and $\mathcal{S}_{i_{l'+1}} \notin N'(\Omega)$; (ii) for all $l \in \{1, 2, \ldots, k\}$, $\mathcal{T}_{j_l} \in \Omega$ and $\mathcal{S}_{i_l} \in N'(\Omega)$; (iii) for all $l \in \{1, 2, \ldots, k\}$, $\mathcal{T}_{j_l} \notin \Omega$; and (iv) there exists some $l^* \in \{1, 2, \ldots, k\}$ such

that $\mathcal{T}_{j_{l^*}} \notin \Omega$ and $\mathcal{S}_{i_{l^*}} \in N'(\Omega)$. These four cases cover all the possibilities. To see this, suppose that none of cases (i–iii) holds. Since neither case (ii) or (iii) holds, there are consecutive l^*-1 and l^* such that $\mathcal{T}_{j_{l^*-1}} \in \Omega$ and $\mathcal{T}_{j_{l^*}} \notin \Omega$. Also, because case (i) does not hold, we must also have $\mathcal{S}_{i_{l^*}} \in N'(\Omega)$, which implies that we are in case (iv).

Case (i). Suppose that there exists some $l' \in \{1,2,\ldots,k\}$ such that $\mathcal{T}_{j_{l'}} \in \Omega$ and $\mathcal{S}_{i_{l'+1}} \notin N'(\Omega)$. Then,

$$\begin{aligned}\sum_{\mathcal{S}_i \in N(\Omega)} c_i - \sum_{\mathcal{T}_j \in \Omega} \lambda'_j &\geq c_{i_{l'+1}} + \sum_{l=1}^{k} \Big(\sum_{\mathcal{S}_i \in N'(\Omega \cap \mathscr{C}_l)} c_i - \sum_{\mathcal{T}_j \in \Omega \cap \mathscr{C}_l} \lambda'_j \Big) \\ &\geq c_{i_{l'+1}} + \sum_{\mathcal{T}_{j_l} \in \Omega} \min\{\Delta_l, 0\} \\ &\geq c_{i_{l'+1}} - (k-1)\delta \\ &\geq \frac{c_{\min}}{k} \geq \delta,\end{aligned}$$

where the second inequality follows from (23-26); the third inequality follows from the fact that $\sum_{\mathcal{T}_{j_l} \in \Omega} \min\{\Delta_l, 0\} + \sum_{\mathcal{T}_{j_l} \notin \Omega} \Delta_l \leq \sum_{l=1}^{k} \Delta_l = 0$ and (22); and the fourth inequality follows from (28).

Case (ii). Suppose that $\mathcal{T}_{j_l} \in \Omega$ and $\mathcal{S}_{i_l} \in N'(\Omega)$ for all $l \in \{1,2,\ldots,k\}$. Then, because $\Omega \subsetneq \{\mathcal{T}_1,\ldots,\mathcal{T}_n\}$, there must exist some l' such that $\Omega \cap \mathscr{C}_{l'} \neq \{\mathcal{T}_{j_{l'-1}+1} \ldots \mathcal{T}_{j_{l'}}\}$. Thus, by (24) and (25),

$$\sum_{\mathcal{S}_i \in N(\Omega)} c_i - \sum_{\mathcal{T}_j \in \Omega} \lambda'_j \geq \sum_{l=1}^{k} \Big(\sum_{\mathcal{S}_i \in N'(\mathscr{C}_l \cap \Omega)} c_i - \sum_{\mathcal{T}_j \in \mathscr{C}_l \cap \Omega} \lambda'_j \Big) \geq \delta + \sum_{l=1}^{k} \Delta_l = \delta.$$

Case (iii). Suppose that for all $l \in \{1,2,\ldots,k\}$, $\mathcal{T}_{j_l} \notin \Omega$. Then, by our assumption that the subgraph induced by Ω is connected under $\mathcal{A}$, and the fact that

$$\mathcal{A} = \mathcal{A}' \cup \{(\mathcal{S}_{i_2}, \mathcal{T}_{j_1}), (\mathcal{S}_{i_3}, \mathcal{T}_{j_2}), \ldots, (\mathcal{S}_m, \mathcal{T}_{j_{k-1}}), (\mathcal{S}_{i_1}, \mathcal{T}_n)\},$$

we must have that $\Omega \subset \mathscr{C}_l$ for some $l \in \{1,2,\ldots,k\}$. Then, by (26), we have

$$\sum_{\mathcal{S}_i \in N(\Omega)} c_i - \sum_{\mathcal{T}_j \in \Omega} \lambda'_j \geq \delta.$$

Case (iv). Suppose that there exists some $l^* \in \{1,2,\ldots,k\}$, such that $\mathcal{T}_{j_{l^*}} \notin \Omega$ and $\mathcal{S}_{i_{l^*}} \in N'(\Omega)$. Under this case, $\mathcal{S}_{i_{l^*}}$ must have more than one neighbors under $\mathcal{A}'$, and because of the way Algorithm 1 constructs $\mathcal{A}'$, we must have $N'(\mathcal{T}_{j_{l^*}}) = \{\mathcal{S}_{i_{l^*}}\}$. This in turn implies that $N'(\Omega \cap \mathscr{C}_{l^*}) = N'(\Omega \cap \mathscr{C}_{l^*} \cup \{\mathcal{T}_{j_{l^*}}\})$. Then, by (25), we have that

$$\sum_{\mathcal{S}_i \in N'(\Omega \cap \mathscr{C}_{l^*} \cup \{\mathcal{T}_{j_{l^*}}\})} c_i - \sum_{\mathcal{T}_j \in \Omega \cap \mathscr{C}_{l^*} \cup \{\mathcal{T}_{j_{l^*}}\}} \lambda'_j \geq \Delta_{l^*},$$

$$\implies \sum_{\mathcal{S}_i \in N'(\Omega \cap \mathscr{C}_{l^*})} c_i - \sum_{\mathcal{T}_j \in \Omega \cap \mathscr{C}_{l^*}} \lambda'_j \geq \Delta_{l^*} + \lambda'_{j_{l^*}}.$$

Thus, we have

$$\begin{aligned}
\sum_{\mathcal{S}_i \in N(\Omega)} c_i - \sum_{\mathcal{T}_j \in \Omega} \lambda'_j &\geq \Delta_{l^*} + \lambda'_{j_{l^*}} + \sum_{l=1, l \neq l^*}^{k} \Big(\sum_{\mathcal{S}_i \in N'(\mathscr{C}_l \cap \Omega)} c_i - \sum_{\mathcal{T}_j \in \mathscr{C}_l \cap \Omega} \lambda'_j \Big) \\
&\geq \lambda'_{j_{l^*}} + \sum_{\mathcal{T}_l \in \Omega, l \neq l^*} \Delta_l + \Delta_{l^*} \\
&\geq \lambda'_{j_{l^*}} - (k-1)\delta \geq \frac{\lambda'_{\min}}{k} \geq \delta,
\end{aligned}$$

where the second inequality holds by (24) and (25), and the third inequality holds by (22), and the fact that $\sum_{l=1}^{k} \Delta_l = 0$. This completes the proof, as we covered all possible cases. □

We now complete the proof of Proposition 3 by showing that $\mathcal{A}$ has exactly $m+n$ arcs.

Proof of Proposition 3. Consider flexibility structure $\mathcal{A}$ defined by Equation (27) in Lemma 3. By Lemma 3, $\mathcal{A}$ has a GCG of at least δ. Because $\mathcal{A}'$ contains no cycles and has k components, it must contain exactly $m+n-k$ arcs. Thus, the number of arcs of $\mathcal{A}$ is exactly $m+n-k+k = m+n$.

Finally, $\mathcal{A}$ can be constructed by first running Algorithm 1 (the partition procedure) to obtain $\mathcal{A}'$, and then implementing Equation (27) (the join procedure). The join procedure requires just k operations, where $k \leq \min\{m,n\}$. The complexity of the partition procedure can be computed by noting that Algorithm 1 terminates whenever the indices $i \geq m$ or $j \geq n$. Because at each iteration, either i, j, or both are increased by one, Algorithm 1 always terminates in $O(m+n)$ operations. Therefore, the algorithmic implementation of the partition and join procedures to generate $\mathcal{A}$ takes at most $O(m+n)$ operations. □

With Proposition 3, we are now ready to complete the proof of Theorem 1, demonstrating that in unbalanced and asymmetric systems, there exists a flexibility structure with $m+n$ arcs that performs as well as (asymptotically) full flexibility.

Proof of Theorem 1. By Proposition 3, we can find a flexibility structure $\mathcal{A}$ with GCG at least $\delta = \min\{\lambda'_{\min}, c_{\min}\}/\min\{m,n\}$. We will now show that $\mathcal{A}$ is the flexibility structure that satisfies Equation (19).

Applying Proposition 1, Corollary 1, and the lower bound on GCG of $\mathcal{A}$, we have that

$$\begin{aligned}
\frac{BL(\mathcal{A})}{BL(\mathcal{F})} &\leq \Big(\frac{\Sigma^2}{2n\zeta} - \frac{C - n\zeta}{2} \Big)^{-1} \cdot \Big(\frac{\Sigma^2}{2n\zeta} + \frac{K_1 + \delta K_2}{\delta\sqrt{\zeta}} \Big) \\
&= 1 + \Big(\frac{C - n\zeta}{2} + \frac{K_1 + \delta K_2}{\delta\sqrt{\zeta}} \Big) \cdot \Big(\frac{\Sigma^2}{2n\zeta} - \frac{C - n\zeta}{2} \Big)^{-1} \\
&\leq 1 + \Big(\frac{C - n\zeta}{2} + \frac{K_1 + \delta K_2}{\delta\sqrt{\zeta}} \Big) \cdot \Big(\frac{\Sigma^2}{2n\zeta} - \frac{C}{2} \Big)^{-1},
\end{aligned} \tag{29}$$

where K_1 and K_2 are constants defined in Proposition 1 that only depends on l and u (the upper and lower bounds in Assumption 1). From Assumption 1, we have $\frac{l^2}{2nu} \le \frac{l^2}{2C} = \frac{2nl^2}{4nC} \le \frac{2\Sigma^2}{4nC}$, implying that if $\zeta \le \frac{l^2}{2nu}$, we must have $\frac{C}{2} \le \frac{\Sigma^2}{4n\zeta}$. Substituting this into (29), we get that if $\zeta \le \frac{l^2}{2nu}$, then

$$\frac{BL(\mathcal{A})}{BL(\mathcal{F})} \le 1 + \Big(\frac{C - n\zeta}{2} + \frac{K_1 + \delta K_2}{\delta\sqrt{\zeta}}\Big) \cdot \Big(\frac{4n\zeta}{\Sigma^2}\Big) \le 1 + \Big(\frac{4n\sqrt{\zeta}}{\Sigma^2}\Big)\Big(\frac{\sqrt{\zeta}C}{2} + \frac{K_1 + \delta K_2}{\delta}\Big).$$

Recall from Assumption 1 that both δ and Σ^2 are lower-bounded by positive constants depending on l. Therefore, we have that

$$\frac{BL(\mathcal{A})}{BL(\mathcal{F})} \le 1 + K\sqrt{\zeta}, \quad \text{if } \zeta \le \frac{l^2}{2nu},$$

where K is a positive constant that only depends on l and u. This completes the proof. □

Discussion. By Proposition 1, we also know that for the flexibility structure $\mathcal{A}$ constructed in Theorem 1, it is also true that

$$\frac{MW(\mathcal{A})}{BL(\mathcal{F})} \le 1 + K\sqrt{\zeta}, \tag{30}$$

where $MW(\mathcal{A})$ is the long-run average backlogging cost under the Max-Weight policy. This is important to practitioners, because the optimal production policy under $\mathcal{A}$ is in general intractable, while the Max-Weight policy is easy to implement and only requires solving a simple linear program at each time period. Therefore, it is interesting to note that Theorem 1 and Equation (30) imply that a little bit of flexibility ($m+n$ arcs) not only gives us the ability to achieve similar performance as full flexibility, but also makes the production scheduling much easier.

4.2. Computing GCG

Proposition 1 and the development of Theorem 1 illustrate that the GCG of a partial flexibility structure $\mathcal{A}$ plays a crucial role to the performance of $\mathcal{A}$ in multi-period systems. Therefore, it is interesting to understand how to compute the GCGs of partial flexibility structures. Next, we show that computing GCG for any given $\mathcal{A}$ can be done efficiently. More specifically, we present an algorithm that solves exactly $n(n-1)$ linear programs, and uses the objective values of the max-flow problems to determine a flexibility structure's GCG. The algorithm is stated in Proposition 4 below.

PROPOSITION 4. *Define for each $k = 1, \ldots, n$, $\ell = 1, \ldots, n$, $k \neq \ell$, the linear program $(P_{k,\ell})$ where*

$$(P_{k,\ell}) \qquad \min \sum_{i=1}^{m} c_i p_i - \sum_{j=1}^{n} \lambda'_j q_j$$

$$\begin{aligned} s.t. \quad & p_i - q_j \geq 0, \quad \forall (\mathcal{S}_i, \mathcal{T}_j) \in \mathcal{A}, \\ & q_k = 1, q_\ell = 0, \\ & 0 \leq p_i, q_j \leq 1, \quad \forall i = 1, \ldots, m, \forall j = 1, \ldots, n. \end{aligned}$$

For a given flexibility structure $\mathcal{A}$, let the optimal objective value of $(P_{k,\ell})$ be $x_{k,\ell}$. Then the GCG of $\mathcal{A}$ is equal to $\min_{1 \leq k,\ell \leq n, k \neq \ell} x_{k,\ell}$.

Proof of Proposition 4. Let η denote the GCG of $\mathcal{A}$. By definition of GCG, it can be rewritten as

$$\eta = \min_{1 \leq k,\ell \leq n, k \neq \ell} \left\{ \min_{\substack{\Omega \subseteq \{\mathcal{T}_1, \ldots, \mathcal{T}_n\} \\ \mathcal{T}_k \in \Omega, \mathcal{T}_l \notin \Omega}} \left\{ \sum_{\mathcal{S}_i \in N(\Omega)} c_i - \sum_{\mathcal{T}_j \in \Omega} \lambda'_j \right\} \right\}.$$

Now, for each pair of (k, ℓ), the following optimization problem

$$\min_{\substack{\Omega \subseteq \{\mathcal{T}_1, \ldots, \mathcal{T}_n\} \\ \mathcal{T}_k \in \Omega, \mathcal{T}_l \notin \Omega}} \left\{ \sum_{\mathcal{S}_i \in N(\Omega)} c_i - \sum_{\mathcal{T}_j \in \Omega} \lambda'_j \right\},$$

can be written as the following integer optimization problem.

$$\begin{aligned} (IP_{k,\ell}) \qquad & \min \sum_{i=1}^{m} c_i p_i - \sum_{j=1}^{n} \lambda'_j q_j \\ \text{s.t.} \quad & p_i - q_j \geq 0, \quad \forall (\mathcal{S}_i, \mathcal{T}_j) \in \mathcal{A}, \\ & q_k = 1, q_\ell = 0, \\ & p_i \in \{0,1\}, q_j \in \{0,1\}, \quad \forall i = 1, \ldots, m, \forall j = 1, \ldots, n. \end{aligned}$$

Finally, note that $(P_{k,\ell})$ is a natural linear programming relaxation of $(IP_{k,l})$ that relaxes the integer constraints on p_i and q_j to non-negativity constraints. Moreover, the constraint set of $(P_{k,\ell})$ is totally unimodular, because the dual of $(P_{k,\ell})$ has only network flow constraints. Hence, η is precisely the minimal of all $x_{k,\ell}$'s, which is obtained by solving exactly $n(n-1)$ linear programs $(P_{k,\ell})$ for each pair of (k, ℓ) with $k, \ell \in \{1, 2, \ldots, n\}$, and $k \neq \ell$. □

5. Necessity of $m+n$ Flexibility Arcs

In §4, we showed that it is always possible to design an efficient flexibility structure using $m+n$ flexibility arcs. More specifically, given any parameter inputs (satisfying Assumption 1), we can construct a flexibility structure with $m+n$ arcs, such that the ratio of its backlogging cost (under the optimal policy) to that of full flexibility is close to 1, when the capacity slack is small. In this section, we show that the requirement of $m+n$ arcs is tight. That is, there exist systems where the ratio of the backlogging cost of any flexibility structure with at most $m+n-1$ arcs, to that

of a fully flexible system, is strictly greater than 1 for any capacity slack. We first establish this tightness for systems with 2 plants and 2 products (which we call 2-by-2 systems) in §5.1. Then, in §5.2, we extend results in §5.1 to establish tightness for more general systems.

5.1. Tightness in 2-by-2 Systems

In this subsection, we will show that there exist 2-by-2 systems with arbitrarily small capacity slack ζ, where the ratio of the performance of any flexibility structure with at most 3 arcs to that of full flexibility is strictly greater than 1. To this end, we first consider the so-called $\mathcal{N}$-system, and derive a lower bound on its performance. We then construct an example 2-by-2 system where any flexibility structure with 3 arcs has sub-optimal performance.

An $\mathcal{N}$-system has 2 plants, 2 products, and 3 flexibility arcs. More specifically, without loss of generality, the flexibility structure $\mathcal{A}$ is given by

$$\mathcal{A} = \{(\mathcal{S}_1, \mathcal{T}_1), (\mathcal{S}_2, \mathcal{T}_2), (\mathcal{S}_1, \mathcal{T}_2)\}. \tag{31}$$

The following proposition gives a lower bound on the performance measure $\Gamma(\pi)$, for any feasible production policy π.

PROPOSITION 5. *Consider an $\mathcal{N}$-system with $c_1 > \lambda_1$ and $c_1 + c_2 > \lambda_1 + \lambda_2$. Then,*

$$BL(\mathcal{N}) \geq \frac{\sigma_1^2}{2(c_1 - \lambda_1)} + \frac{\sigma_2^2}{2(c_1 + c_2 - \lambda_1 - \lambda_2)} - \frac{2\lambda_1 + \lambda_2}{2}. \tag{32}$$

The proof of Proposition 5 consists of two parts. We first show that a priority-based production policy is optimal for the $\mathcal{N}$-system, and then derive a performance bound under this priority policy using drift analysis.

LEMMA 4. *Consider the $\mathcal{N}$-system with T finite time periods. Let $J^t(b_1, b_2)$ be the total expected backlog under the optimal policy from period t to T, given that the backlog at products 1 and 2 are b_1 and b_2 before demand arrives at time t. Then, for any $1 \leq t \leq T$, and any $b_1 \geq 1$, $b_2 \geq 0$, we have*

$$J^t(b_1, b_2) \geq J^t(b_1 - 1, b_2 + 1).$$

Proof of Lemma 4 is in §EC.4. The following corollary is immediate from Lemma 4.

COROLLARY 3. *Consider the policy π^*, that always first uses plant 1 to satisfy the demand and backlog of product 1, and then uses the leftover capacity at plant 1 and capacity at plant 2 to satisfy the demand and backlog of product 2. Then, for any $T \geq 0$, π^* is a policy that minimizes*

$$\min_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} (B_1^{\pi}(t) + B_2^{\pi}(t))\right], \tag{33}$$

among any feasible policy π.

In the interest of notation simplicity, for the rest of this section, we use $B_1(t)$ and $B_2(t)$ to denote the backlog of products 1 and 2 under the optimal policy π^*. Therefore, the backlogs of products 1 and 2 evolve as follows:

$$B_1(t) = B_1(t-1) + D_1(t) - c_1 + U_1(t), \tag{34}$$

$$B_2(t) = B_2(t-1) + D_2(t) - c_2 - U_1(t) + U_2(t), \tag{35}$$

where $U_1(t)$ represents the leftover capacity at plant 1 at time t, after producing product 1, and is defined to be

$$U_1(t) = \Big(c_1 - D_1(t) - B_1(t-1)\Big)^+, \tag{36}$$

and $U_2(t)$ represents the unused capacities at time t from both plants 1 and 2 after all production decisions have been made, and is defined to be

$$U_2(t) = \Big(c_2 + U_1(t) - D_2(t) - B_2(t-1)\Big)^+. \tag{37}$$

When $c_1 > \lambda_1$ and $c_1 + c_2 > \lambda_1 + \lambda_2$, the backlog process in the $\mathcal{N}$-system under policy π^* is stable and converges to a unique equilibrium distribution. We use (the law of) the random vector $(B_1(\infty), B_2(\infty))$ to represent this distribution. As a result, the long-run average backlog (under the optimal policy π^*) can be expressed as:

$$BL(\mathcal{N}) = \lim_{T\to\infty} \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} (B_1(t) + B_2(t))\right] = \mathbb{E}[B_1(\infty) + B_2(\infty)].$$

To lower bound $BL(\mathcal{N})$, we derive lower bounds on $\mathbb{E}[B_1(\infty)]$ and $\mathbb{E}[B_2(\infty)]$ respectively.

First, note that $B_1(t)$ is simply the backlog at time t of a single product, single plant production system. Thus, we have the following lower bound on $\mathbb{E}[B_1(\infty)]$ from Proposition 2:

$$\mathbb{E}[B_1(\infty)] \geq \frac{\sigma_1^2}{2(c_1 - \lambda_1)} - \frac{\lambda_1}{2}. \tag{38}$$

Next, we derive a lower bound on $\mathbb{E}[B_2(\infty)]$, by analyzing the expected drift of $B_2^2(t)$ in steady state. The analysis uses the negative correlation between $B_2(t-1)$ and $U_1(t)$ in a crucial way, which we establish below, by first invoking the following simple corollary of a result from Müller and Stoyan (2002).

LEMMA 5. *Let $X_1,\ldots,X_n$ be a sequence of independent random variables. Let $f(x_1,\ldots,x_n)$ and $g(x_1,\ldots,x_n)$ be functions that are non-decreasing in each component x_k. Then, $f(X_1,\ldots X_n)$ and $g(X_1,\ldots,X_n)$ are positively correlated, i.e.,*

$$\mathbb{E}[f(X_1,\ldots X_n)g(X_1,\ldots,X_n)] \geq \mathbb{E}[f(X_1,\ldots X_n)]\mathbb{E}[g(X_1,\ldots,X_n)].$$

Proof of Lemma 5. Applying Theorem 3.10.5 (v) from Müller and Stoyan (2002), we have that the random vector $\mathbf{X} = (X_1,\ldots,X_n)$ is associated, which implies that

$$\mathbb{E}[f(X_1,\ldots X_n)g(X_1,\ldots,X_n)] \geq \mathbb{E}[f(X_1,\ldots X_n)]\mathbb{E}[g(X_1,\ldots,X_n)]. \quad \square$$

LEMMA 6. *Suppose that the system is empty initially, i.e., $B_1(0) = B_2(0) = 0$. Then, for any $t \geq 1$,*

$$\mathbb{E}[B_2(t-1)U_1(t)] \leq \mathbb{E}[B_2(t-1)]\mathbb{E}[U_1(t)]. \tag{39}$$

Proof of Lemma 6. We first show that $B_2(t-1)$ and $-U_1(t)$ can be expressed as non-decreasing functions applied to $D_1(1),\ldots,D_1(t)$, $D_2(1),\ldots,D_2(t)$.

Let $\mathbf{d} = [d_1(1),\ldots,d_1(t),\ d_2(1),\ldots,d_2(t)]$ be an arbitrary sequence of demand realizations for products 1 and 2. Under demand realization $\mathbf{d}$, by Equations (34) and (35), $b_1(x)$ and $b_2(x)$, the backlogs for product 1 and 2 at time period x, for $1 \leq x \leq t$, can be determined iteratively as

$$b_1(x) = (b_1(x-1) + d_1(x) - c_1)^+, \tag{40}$$

$$b_2(x) = \Big(b_2(x-1) + d_2(x) - c_2 - (c_1 - d_1(x) - b_1(x-1))^+\Big)^+, \tag{41}$$

with $b_1(0) = b_2(0) = 0$.

Both $b_1(x)$ and $b_2(x)$ can be viewed as a function that map $\mathbf{d}$ to a real number. For some $1 \leq x \leq t$, suppose that $b_1(x-1)$ and $b_2(x-1)$ are non-decreasing in $\mathbf{d}$. Then, by Equation (40), $b_1(x)$ is non-decreasing in $b_1(x-1)$ and $d_1(x)$, which implies that $b_1(x)$ is non-decreasing in $\mathbf{d}$. By Equation (41), $b_2(x)$ is non-decreasing in $b_2(x-1)$, $d_2(x)$, $d_1(x)$ and $b_1(x-1)$, so $b_2(x)$ is also non-decreasing in $\mathbf{d}$. Because $b_1(0) = b_2(0) = 0$, by induction, we have that $b_1(t-1)$ and $b_2(t-1)$ are non-decreasing in $\mathbf{d}$. Finally, by Equation (36),

$$u_1(t) = \Big(c_1 - d_1(t) - b_1(t-1)\Big)^+,$$

implying that $u_1(t)$ is decreasing in both $b_1(t-1)$ and $d_1(t)$ and therefore in $\mathbf{d}$.

Therefore, $B_2(t-1)$ and $-U_1(t)$ can be rewritten as

$$B_2(t-1) = f(D_1(1),\ldots,D_1(t),D_2(1),\ldots,D_2(t))$$

$$-U_1(t) = g(D_1(1), \ldots, D_1(t), D_2(1), \ldots, D_2(t)),$$

where f and g are non-decreasing functions. By Lemma 5, and the fact that $D_1(1), \ldots, D_1(t), D_2(1), \ldots, D_2(t)$ are all independent, we get

$$-\mathbb{E}[B_2(t-1)U_1(t)] \geq -\mathbb{E}[B_2(t-1)]\mathbb{E}[U_1(t)], \text{ i.e., } \mathbb{E}[B_2(t-1)U_1(t)] \leq \mathbb{E}[B_2(t-1)]\mathbb{E}[U_1(t)]. \quad \square$$

We are now ready to derive the lower bound on $\mathbb{E}[B_2(\infty)]$.

LEMMA 7. *Consider an $\mathcal{N}$-system with $c_1 > \lambda_1$ and $c_1 + c_2 > \lambda_1 + \lambda_2$. Recall that $B_2(\infty)$ denotes the steady state equilibrium backlog of product 2 under policy π^* described in Corollary 3. Then, we have*

$$\mathbb{E}[B_2(\infty)] \geq \frac{\sigma_2^2}{2(c_1 + c_2 - \lambda_1 - \lambda_2)} - \frac{\lambda_1 + \lambda_2}{2}. \tag{42}$$

Proof of Lemma 7. We first note the following useful identity: $U_2(t)B_2(t) = 0$. This is true since if $U_2(t) > 0$, i.e., there is positive unused capacity, then the backlog in the next time period must have been cleared, i.e., $B_2(t) = 0$. Consequently, $U_2(t)(B_2(t-1) + D_2(t) - c_2 - U_1(t) + U_2(t)) = 0$ for all t. This also implies that $\mathbb{E}[U_2(t)(B_2(t-1) + D_2(t) - c_2 - U_1(t))] = -\mathbb{E}[U_2^2(t)]$.

Other two useful identities are $\mathbb{E}[U_1(\infty)] = c_1 - \lambda_1$ and $\mathbb{E}[U_2(\infty)] = c_1 + c_2 - \lambda_1 - \lambda_2$, where $(U_1(\infty), U_2(\infty))$ has the limiting distribution of $(U_1(t), U_2(t))$. To see that $\mathbb{E}[U_1(\infty)] = c_1 - \lambda_1$, take expectation on both sides of Equation (34), and let $t \to \infty$. Then, we must have $\mathbb{E}[D_1 - c_1 + U_1(\infty)] = 0$. Since $\mathbb{E}[D_1] = \lambda_1$, we have $\mathbb{E}[U_1(\infty)] = c_1 - \lambda_1$. $\mathbb{E}[U_2(\infty)] = c_1 + c_2 - \lambda_1 - \lambda_2$ can be derived similarly by taking expectation on both sides of Equation (35) and letting $t \to \infty$.

Next, consider the change in the second moment of B_2, i.e., $\mathbb{E}[B_2^2(t) - B_2^2(t-1)]$. We have

$$\begin{aligned}
&\mathbb{E}[B_2^2(t) - B_2^2(t-1)] \\
&= \mathbb{E}[(B_2(t-1) + D_2(t) - c_2 - U_1(t) + U_2(t))^2 - B_2^2(t-1)] \\
&= \mathbb{E}[(B_2(t-1) + D_2(t) - c_2 - U_1(t) + U_2(t))(B_2(t-1) + D_2(t) - c_2 - U_1(t)) - B_2^2(t-1)] \\
&= \mathbb{E}[(B_2(t-1) + D_2(t) - c_2 - U_1(t))^2 + U_2(t)(B_2(t-1) + D_2(t) - c_2 - U_1(t)) - B_2^2(t-1)] \\
&= \mathbb{E}[(B_2(t-1) + D_2(t) - c_2 - U_1(t))^2 - U_2^2(t) - B_2^2(t-1)] \\
&= \mathbb{E}[(D_2(t) - c_2 - U_1(t))^2] + 2\mathbb{E}[B_2(t-1) \cdot (D_2(t) - c_2 - U_1(t))] - \mathbb{E}[U_2^2(t)] \\
&\geq \mathbb{E}[(D_2(t) - c_2 - U_1(t))^2] + 2\mathbb{E}[B_2(t-1)]\mathbb{E}[D_2(t) - c_2 - U_1(t)] - \mathbb{E}[U_2^2(t)],
\end{aligned}$$

where the last inequality follows from Lemma 6 and the fact that $B_2(t-1)$ is independent from $D_2(t)$.

Now, letting t go to infinity, and noting that in steady state, the expected drift of $B_2^2(t)$ is 0, we have

$$0 \geq \mathbb{E}[(D_2 - c_2 - U_1(\infty))^2] - 2\mathbb{E}[B_2(\infty)](c_1 + c_2 - \lambda_1 - \lambda_2) - \mathbb{E}[U_2^2(\infty)],$$

Rearranging the inequality, we get that

$$\mathbb{E}[B_2(\infty)] \geq \frac{\mathbb{E}[(D_2 - c_2 - U_1(\infty))^2]}{2(c_1 + c_2 - \lambda_1 - \lambda_2)} - \frac{\mathbb{E}[U_2^2(\infty)]}{2(c_1 + c_2 - \lambda_1 - \lambda_2)}. \tag{43}$$

Furthermore, we can lower-bound $\mathbb{E}[(D_2 - c_2 - U_1(\infty))^2]$ as

$$\begin{aligned}
&\mathbb{E}[(D_2 - c_2 - U_1(\infty))^2] \\
&= \mathbb{E}[(D_2 - c_2)^2] - 2\mathbb{E}[D_2 - c_2]\mathbb{E}[U_1(\infty)] + \mathbb{E}[U_1^2(\infty)] \\
&\geq \mathbb{E}[(D_2 - c_2)^2] - 2\mathbb{E}[D_2 - c_2]\mathbb{E}[U_1(\infty)] + \mathbb{E}[U_1(\infty)]^2 \\
&= (c_2 - \lambda_2)^2 + \sigma_2^2 + 2(c_2 - \lambda_2)(c_1 - \lambda_1) + (c_1 - \lambda_1)^2 \\
&= (c_1 + c_2 - \lambda_1 - \lambda_2)^2 + \sigma_2^2;
\end{aligned} \tag{44}$$

and upper-bound $\mathbb{E}[U_2^2(\infty)]$ as

$$\mathbb{E}[U_2^2(\infty)] \leq (c_1 + c_2)\mathbb{E}[U_2(\infty)] = (c_1 + c_2)(c_1 + c_2 - \lambda_1 - \lambda_2). \tag{45}$$

Substituting Inequalities (44) and (45) into Inequality (43), we have

$$\mathbb{E}[B_2(\infty)] \geq \frac{\sigma_2^2}{2(c_1 + c_2 - \lambda_1 - \lambda_2)} - \frac{\lambda_1 + \lambda_2}{2}. \quad \square \tag{46}$$

Proof of Proposition 5. We have now obtained lower-bounds on both $\mathbb{E}[B_1(\infty)]$ and $\mathbb{E}[B_2(\infty)]$. The proof of Proposition 5 can be concluded by adding the lower-bounds in Equations (38) and (42). $\square$

A Counter-example. We next apply Proposition 5 to show that 3 flexibility arcs in a 2-by-2 system may not achieve the same asymptotic as full flexibility. Consider a system with 2 plants, 2 products, $\lambda_1 = 1 - 3\varepsilon$, $\lambda_2 = 1 - \varepsilon$, and $c_1 = c_2 = 1$, where $\varepsilon > 0$ can be thought of as an arbitrarily small constant. For concreteness, suppose that $\sigma_1^2 = \sigma_2^2 = 1$, although we only require σ_1^2 and σ_2^2 to be fixed positive constants. Then, the capacity slack $\zeta = 2\varepsilon$. Furthermore, $\lambda_1' = 1 - \varepsilon$ and $\lambda_2' = 1 + \varepsilon$. It is not difficult to see that the only flexibility structure with 3 arcs that has a positive GCG is the $\mathcal{N}$-structure given by (31), where the GCG $\eta = \varepsilon$. By Proposition 5,

$$BL(\mathcal{N}) \geq \frac{\sigma_1^2}{6\varepsilon} + \frac{\sigma_2^2}{8\varepsilon} - 1.5 = \frac{7}{24\varepsilon} - 1.5.$$

In contrast, under the fully flexible structure, by Proposition 2,

$$BL(\mathcal{F}) \le \frac{\sigma_1^2+\sigma_2^2}{8\varepsilon} + 2\varepsilon = \frac{1}{4\varepsilon} + 2\varepsilon.$$

Thus,

$$\liminf_{\varepsilon\to 0} \frac{BL(\mathcal{N})}{BL(\mathcal{F})} > 1.$$

Let us note that our counter-example illustrates the following important point. For each $\varepsilon > 0$, the corresponding $\mathcal{N}$-system satisfies $\eta > 0$, which is equivalent to the CRP condition discussed in §3.1. But for any arbitrarily small ζ (this can be obtained by making ε small), the system performance is at least a constant factor away from that of the fully flexible system. Thus, to design effective flexible structures, we cannot rely on the CRP condition alone.

5.2. Tightness in General Systems

The need for $m+n$ arcs in a flexibility structure to achieve near-optimal performance is also necessary in systems that are more general than those with 2 plants and 2 products. We construct a suite of counterexamples in this section. We do want to caution the readers that the requirement of $m+n$ arcs is not necessary for all systems. For example, the full flexiblility structure for $m=1$ and $n=10$ contains only $m+n-1$ arcs.

Similar to §5.1, we first derive a lower bound on $BL(\mathcal{A})$, the performance of any given flexibility structure $\mathcal{A}$, using a simple coupling argument.

PROPOSITION 6. *Consider a general system with m plants, n products, capacity vector $\mathbf{c}$, demand rate vector $\boldsymbol{\lambda}$, demand variance vector $\boldsymbol{\sigma}$, and a connected flexibility structure $\mathcal{A}$ with a positive GCG $\eta > 0$. Suppose that GCG is attained at the set Ω, so that $\eta = \sum_{\mathcal{S}_i \in N(\Omega)} c_i - \sum_{\mathcal{T}_j \in \Omega} \lambda'_j$. Let $\Sigma_\Omega^2 = \sum_{\mathcal{T}_j\in\Omega} \sigma_j^2$, and $\Sigma_{\Omega^c}^2 = \sum_{\mathcal{T}_j\notin\Omega} \sigma_j^2$. Then,*

$$BL(\mathcal{A}) \ge \frac{\Sigma_\Omega^2}{2(\eta+|\Omega|\zeta)} + \frac{\Sigma_{\Omega^c}^2}{2n\zeta} - \frac{\Lambda + \sum_{\mathcal{T}_j\in\Omega}\lambda_j}{2}. \tag{47}$$

Proof of Proposition 6. Consider the coupling of the system of interest with an $\mathcal{N}$-system described as follows. In this $\mathcal{N}$-system, the demand for product 1 at time t is given by $\tilde{D}_1(t) = \sum_{\mathcal{T}_j\in\Omega} D_j(t)$, demand for product 2 at time t is given by $\tilde{D}_2(t) = \sum_{\mathcal{T}_j\notin\Omega} D_j(t)$, capacity at plant 1 is given by $\tilde{c}_1 = \sum_{\mathcal{S}_i\in N(\Omega)} c_i$, and $\tilde{c}_2 = \sum_{\mathcal{S}_i\notin N(\Omega)} c_i$. Let the backlogs for products 1 and 2 at time t be denoted by $\tilde{B}_1(t)$ and $\tilde{B}_2(t)$, respectively. If the initial backlogs in the original system are given by $\{B_j(0)\}$, then the initial backlogs of the $\mathcal{N}$-system are given by

$$\tilde{B}_1(0) = \sum_{\mathcal{S}_j\in\Omega} B_j(0); \qquad \tilde{B}_2(0) = \sum_{\mathcal{S}_j\notin\Omega} B_j(0).$$

It is clear that $\mathbb{E}[\tilde{D}_1(t)] = \sum_{\mathcal{T}_j\in\Omega}\lambda_j$, $\mathbb{E}[\tilde{D}_2(t)] = \sum_{\mathcal{T}_j\notin\Omega}\lambda_j$, $Var(\tilde{D}_1(t)) = \sum_{\mathcal{T}_j\in\Omega}\sigma_j^2 = \Sigma_\Omega^2$, and $Var(\tilde{D}_2(t)) = \sum_{\mathcal{T}_j\notin\Omega}\sigma_j^2 = \Sigma_{\Omega^c}^2$. Furthermore, $\tilde{c}_1 - \sum_{\mathcal{S}_j\in\Omega}\lambda_j = \eta + |\Omega|\zeta$, and $\tilde{c}_1 + \tilde{c}_2 - \sum_j \lambda_j = n\zeta$.

Consider any production policy π for the original system. It induces a production policy $\tilde{\pi}$ for the $\mathcal{N}$-system, which we define below. Suppose that under π, at time t, the production amount of plant i for product j is given by $f_{i,j}$. Then, define $\tilde{f}_{\tilde{i},\tilde{j}}$, the production amount of plant $\tilde{i}$ for product $\tilde{j}$ under $\tilde{\pi}$ at time t in the $\mathcal{N}$-system to be

$$\tilde{f}_{1,1} = \sum_{\mathcal{S}_i\in N(\Omega),\mathcal{T}_j\in\Omega} f_{i,j}, \qquad \tilde{f}_{1,2} = \sum_{\mathcal{S}_i\in N(\Omega),\mathcal{T}_j\notin\Omega} f_{i,j}, \qquad \tilde{f}_{2,2} = \sum_{\mathcal{S}_i\notin N(\Omega),\mathcal{T}_j\notin\Omega} f_{i,j}.$$

By a simple induction, we can show that with probability 1, $\sum_{j=1}^n B_j(t) \geq \tilde{B}_1(t) + \tilde{B}_2(t)$ for all t. We can now apply Proposition 5 to conclude that

$$\Gamma(\pi) \geq \Gamma(\tilde{\pi}) \geq \frac{\Sigma_\Omega^2}{2(\eta+|\Omega|\zeta)} + \frac{\Sigma_{\Omega^c}^2}{2n\zeta} - \frac{\Lambda + \sum_{\mathcal{T}_j\in\Omega}\lambda_j}{2}. \quad \square$$

We note that the proof Proposition 6 implies a slightly more general statement. In particular, for any set Ω' and any $\eta' = \sum_{\mathcal{S}_i\in N(\Omega')} c_i - \sum_{\mathcal{T}_j\in\Omega'}\lambda'_j$, where η' is not necessarily equal to the GCG, we can obtain a lower bound on $BL(\mathcal{A})$ via Equation (47) by modifying the parameters accordingly. An immediate consequence of Proposition 6 is the following corollary.

COROLLARY 4. *Consider a system with the average slack* $\zeta \leq \min\left\{\frac{l}{n}, \frac{\alpha l^2}{4mn(n-\alpha)u}\right\}$ *and a flexibility structure* $\mathcal{A}$ *such that the GCG* $\eta < (1-\alpha)\zeta$*, for some* $\alpha \in (0,1)$*. Then,*

$$\frac{BL(\mathcal{A})}{BL(\mathcal{F})} \geq 1 + \frac{\alpha l}{4n(n-\alpha)u^2} > 1. \tag{48}$$

The proof of Corollary 4 can be found in §EC.4. From Corollary 4, we see that for any structure $\mathcal{A}$, the magnitude of the GCG η (relative to that of the slack ζ) is crucial in determining the performance of $\mathcal{A}$. We now proceed to provide examples of general sized systems with $m+n-1$ arcs and positive GCG, but whose asymptotic performance is strictly worse than that of full flexibility, generalizing the examples in §5.1.

Counter-examples. For any $m, n \geq 2$, consider a system with m plants and n products. Plants have integral capacities with the total capacity C equal to n, and the demand rates are given by $\lambda_1 = \lambda_2 = \cdots = \lambda_{n-1} = 1-2\varepsilon-\varepsilon/n$, and $\lambda_n = 1-2\varepsilon+\frac{n-1}{n}\varepsilon$. Then, we have $\lambda'_1 = \cdots = \lambda'_{n-1} = 1-\varepsilon/n$, $\lambda'_n = 1+\frac{n-1}{n}\varepsilon$, and $\zeta = 2\varepsilon$. For simplicity, again suppose that $\sigma_j^2 = 1$ for all j.

Observe that we can always construct a structure that has $m+n-1$ arcs and has positive GCG using greedy algorithm. Consider some $\varepsilon < 1$ and suppose that $\mathcal{A}$ is a structure with $m+n-1$ arcs

and positive GCG. For any $1 \le i \le n$, let $\Omega_i^o = \{\mathcal{T}_j | N(\mathcal{T}_j) = \{\mathcal{S}_i\}\}$, i.e., the set of all product nodes that only has $\mathcal{S}_i$ as its neighbor. Then, the total number of arcs of $\mathcal{A}$ is equal to or greater than

$$\sum_{i=1}^{m} |\Omega_i^o| + 2(n - \sum_{i=1}^{m} |\Omega_i^o|) = n + (\sum_{i=1}^{m} c_i - \sum_{i=1}^{m} |\Omega_i^o|) = n + \sum_{i=1}^{m} (c_i - |\Omega_i^o|).$$

Because we assume that $\mathcal{A}$ has $m+n-1$ arcs, we must have some i^* such that $c_{i^*} \le |\Omega_{i^*}^o|$. Now, observe that

$$\Lambda + \sum_{\mathcal{T}_j \in \Omega} \lambda_j \le 2n \tag{49}$$

Also, note that because we assume $\varepsilon < 1$ and $\mathcal{A}$ has positive GCG, we must have $c_{i^*} + 1 > |\Omega_{i^*}^o|$, implying $c_{i^*} = |\Omega_{i^*}^o|$. For notational simplicity, we use to c^* denote $|\Omega_{i^*}^o|$. Combining Equation (49) and Proposition 6, we get

$$\begin{aligned} BL(\mathcal{A}) &\ge \frac{c^*}{2(\varepsilon c^*/n + 2\varepsilon c^*)} + \frac{n - c^*}{2(2n\varepsilon)} - n \\ &= \frac{1}{4\varepsilon} - n + \frac{c^*}{2\varepsilon(c^*/n + 2c^*)} - \frac{c^*}{2(2n\varepsilon)} \\ &> \frac{1}{4\varepsilon} - n + \frac{c^*}{2\varepsilon}\left(\frac{1}{2n-1} - \frac{1}{2n}\right) \\ &= \frac{1}{4\varepsilon} - n + \frac{c^*}{4n(2n-1)\varepsilon}, \end{aligned}$$

where the inequality follows from observing that $c^* \le 2n-1$ thereby $c^*/n + 2c^* < 2n - 1$. The performance of full flexibility is upper-bounded as

$$BL(\mathcal{F}) \le \frac{n}{2n(2\varepsilon)} + \frac{2n(2\varepsilon)}{2} = \frac{1}{4\varepsilon} + 2n\varepsilon.$$

Thus,

$$\liminf_{\varepsilon \to 0} \frac{BL(\mathcal{A})}{BL(\mathcal{F})} = 1 + \frac{c^*}{n(2n-1)} > 1.$$

6. Numerical Experiments

Our theoretical analysis has given us the following insights under highly utilized systems: (i) we can find a sparse flexibility structure with $m+n$ arcs that would perform close to full flexibility; (ii) in some systems, we cannot find a flexibility structure with $m+n-1$ arcs that has a good GCG, implying that it would perform drastically worse compared to a well-designed structure with $m+n$ arcs.

Motivated by the above theoretical findings, we carry out extensive numerical experiments to study the empirical performance of various process flexibility structures. The goal of our simulation study is to understand how insights change as we deviate away from the theoretical assumption

that utilization rate goes to 1, and how robust the insights are when the system size and the variability of product demands change. Besides the discrete-time backlogging environment studied in this paper, we also investigate how our insights extend to continuous-time environments, such as parallel queueing networks, and serial production lines.

6.1. Discrete-Time Backlogging Environments

6.1.1. Balanced and Symmetric Systems. We investigate the performance of long chain/chaining and compare it to other structures in balanced systems. The testing parameters include system sizes ($m = n = 5, 10, 15, 20$), coefficient of variations of demand (cv $= 0.3, 0.4, 0.5$) and utilization rates ($\rho = 0.8, 0.9, 0.95, 0.975, 0.9875$). For each triplet of parameters, we simulate the expected backlogs of dedicated flexibility, long chain, and full flexibility (see examples in Figure 1 for $m = n = 4$). In our simulation, we set the capacity for each plant to be 100, and use independent normal distributions (truncated below at 0 and above at twice of the average) to simulate product demands. The mean demand for product 2 to $n-1$ is set to be 100ρ, and we slightly perturb the means for product 1 and product n to be 95ρ and 105ρ, respectively. The reason for perturbation is to avoid overly optimistic performance of the long chain due to perfect symmetry. (This concern is probably overly cautious, since we have not observed significant differences between the perturbed and completely symmetric systems in numerical simulations.)

We next describe the policies used to evaluate the expected backlogs for different structures. We say that a policy is a Max-Flow policy if, during each period, it finds a production schedule to greedily minimize the total backlog. For dedicated and full flexibility structures, implementing a Max-Flow policy is straightforward and optimal. However, computing the optimal policy under the long chain structure is much more difficult, as it requires solving an infinite horizon dynamic program (DP) where the size of the state space increases exponentially with the system size. To avoid the curse of dimensionality of DP, we compute its expected backlogs under the Max-Weight (MW) policy, motivated from the asymptotically optimal analysis of MW from §3.2. While not all systems we simulate have close to 100% utilization rate, we observe in numerical experiments that for the long chain, MW is always better than other simple heuristics (e.g., a priority policy to be discussed later). Finally, while not optimal, MW under the long chain often exhibits strong performance when benchmarked against dedicated and full flexibility.

For each system, we first run 250 warm-up periods, and then record the average backlogs for the next 50 periods, for 10000 randomly generated samples. In our computational experiments, we do not observe significant differences in our performance measure when we perturb the number of warm-up periods and the number of periods to record backlogs. We use $B(\mathcal{D})$, $B(\mathcal{LC})$, and $B(\mathcal{F})$ to denote the empirical expected average backlogs under dedicated, long chain and full flexibility,

respectively. In the simulations, the expected average backlogs for all flexibility structures under most settings have a standard error within 1% of the empirical expected average backlogs. We use SE%($\mathcal{A}$) to denote the ratio between the standard error of $B(\mathcal{A})$ and $B(\mathcal{A})$ in percentages, and set SE% to be the maximum of SE%($\mathcal{D}$), SE%($\mathcal{LC}$) and SE%($\mathcal{F}$). For the settings with SE% greater than 1%, the expected average backlogs for both the long chain and full flexibility are very close to zero (see Tables EC.1 and EC.2).

To understand the effectiveness of the long chain, we compute two performance measures:

$$R(\mathcal{A}) = \frac{B(\mathcal{A})}{B(\mathcal{F})}, \quad \Delta(\mathcal{A}) = \frac{B(\mathcal{D}) - B(\mathcal{A})}{B(\mathcal{D}) - B(\mathcal{F})}, \quad \text{for any flexibility structure } \mathcal{A}. \tag{50}$$

In particular, $R(\mathcal{LC})$ represents the ratio between the backlogs of long chain and that of full flexibility, and $\Delta(\mathcal{LC})$ represents the ratio between the improvement (starting from dedicated flexibility) of long chain and that of full flexibility. From our theoretical results in §3.2, we know that $R(\mathcal{LC})$ (and thus $\Delta(\mathcal{LC})$) approaches 1 as ρ (utilization rate) goes to 1. Because the backlogs of full flexibility is always less than that of the long chain, $R(\mathcal{LC})$ is greater than 1, while $\Delta(\mathcal{LC})$ is less than 1. The reason we include $\Delta(\mathcal{LC})$ in addition to $R(\mathcal{LC})$ is because in some settings, the expected backlog of full flexibility is extremely close to zero, which may cause $R(\mathcal{LC})$ to be large, and $\Delta(\mathcal{LC})$ seems to be a better measure on the effectiveness of the long chain in this case.

In Tables EC.1 and EC.2, we present $R(\mathcal{LC})$, $\Delta(\mathcal{LC})$, and $B(\mathcal{LC})$, under different parameter settings. These tables help us better understand what happens when ρ is not near 1. For example, when $\rho = 0.8$, in most settings, $R(\mathcal{LC})$ is no longer close to 1. However, this does not suggest that the long chain performs poorly, because the expected backlogs of $R(\mathcal{LC})$ and full flexibility when $\rho = 0.8$ are often very close to zero. A better measure in this case is $\Delta(\mathcal{LC})$, which is at least 97% for all settings with $\rho = 0.8$, implying that the long chain is already capturing at least 97% of the improvement carried by full flexibility. Moreover, when $\rho = 0.8$, the percentage of backlogs under the long chain is also very small, as it never exceeds 2% of the average demand per period. Overall, in all settings, the long chain performs very well when measured using $\Delta(\mathcal{LC})$, as $\Delta(\mathcal{LC})$ never falls below 92%. Therefore, we conclude that in balanced and symmetric systems with $n \leq 20$, while the ratio between the backlogs of the long chain to that of the full flexibility is not necessarily close to 1, the long chain is always very effective based on $\Delta(\mathcal{LC})$. We also find that $\Delta(\mathcal{LC})$ in general decreases as n increases, while keeping all other parameters constant. Therefore, a caveat is that if n is much larger than 20, we should not expect $\Delta(\mathcal{LC})$ to be close to 1 for non-asymptotic utilization rates. This is intuitive, because when n is large, there is a huge difference in the number of flexibility arcs between long chain and full flexibility.

We also simulate the expected backlogs of the long chain *less* the arc $(n, 1)$, which is denoted by $\mathcal{LC}^-$. Simulation results are reported in Table EC.3. The reason we study $\mathcal{LC}^-$ is that it has $2n - 1$

arcs, just one less arc compared to $\mathcal{LC}$, and yet has a much smaller GCG. The theoretical analysis in §5 suggests that a structure with small GCG can have significantly higher backlogs than long chain, and we use simulation to verify this insight. The fact that $\mathcal{LC}^-$ is significantly worse than $\mathcal{LC}$ is also tightly related to the idea of "closing the chain", which has been well established under many different environments by classical literature in process flexibility (see Jordan and Graves (1995), Hopp et al. (2004), and Iravani et al. (2005)).

Similar to the long chain, the optimal policy for $\mathcal{LC}^-$ is also difficult to compute, due to the curse of dimensionality. We evaluate the expected backlogs of $\mathcal{LC}^-$ under each parameter set by picking the better of MW and a priority policy, which is a Max-Flow policy that prioritizes products from the smallest label to the largest (see §EC.3 for implementation details). The reason we do not use MW solely to evaluate the expected backlogs of $\mathcal{LC}^-$ is that $\mathcal{LC}^-$ has a very small GCG, so MW can be significantly sub-optimal especially when ρ is not close to 1. Thus, to ensure that $\mathcal{LC}^-$ is not penalized because of the sub-optimality of MW, we include the priority policy which often performs much better in simulations, to keep our simulation results more robust.

In Table EC.3, we list $R(\mathcal{LC}^-)$ and $\Delta(\mathcal{LC}^-)$ under the same parameter combinations used to test $\mathcal{LC}$. In the interest of space, we omit SE%$(\mathcal{LC}^-)$ as all values are less than the SE% values reported in Tables EC.1 and EC.2. Comparing these numbers with the numbers in Tables EC.1 and EC.2, we see that $R(\mathcal{LC}^-)$ is significantly higher than $R(\mathcal{LC})$, while $\Delta(\mathcal{LC}^-)$ is significantly lower than $\Delta(\mathcal{LC})$. This observation not only matches our theoretical result when utilization rate approaches 1, but also shows that $\mathcal{LC}$ is significantly better than $\mathcal{LC}^-$ when the utilization rate is in the 80% to 90% range. Indeed, this observation echoes the idea of "closing the chain" that has been established in single-period system (Jordan and Graves (1995)), serial production line (Hopp et al. (2004)), and call center (Iravani et al. (2005)). In the next subsection, in unbalanced and asymmetric systems, we also observe that a structure with $m+n$ arcs (with a high GCG) can considerably outperform structures with $m+n-1$ arcs.

6.1.2. Backlogging Environment under Non-Balanced and Asymmetric Systems. Next, we study the performance of sparse structures with $m+n$ and $m+n-1$ arcs in non-balanced and asymmetric systems. We vary system sizes $(m,n)=(3,5),(6,10),(9,15)$, coefficient of variations for demand distribution cv $=0.3,0.4,0.5$, and utilization rates $\rho=0.8,0.9,0.95,0.975,0.9875$. For the system with 3 plants $(m=3)$ and 5 products $(n=5)$, the capacity of each plant is set to be 100, and the vector for mean product demands is set to be $[55\rho,50\rho,50\rho,50\rho,95\rho]$. Like §6.1.1, the distributions of the product demands are set to be independent (truncated) normals. Systems with $(m,n)=(6,10)$ and $(m,n)=(9,15)$ contain two and three copies of the parameters of the 3 by 5 system, respectively.

For each set of system parameters, we study four different structures, namely, full flexibility, a structure with n arcs, a structure with $m+n-1$ arcs, and a structure with $m+n$ arcs. The structures with n, $m+n-1$ and $m+n$ arcs are displayed in Figure EC.2 (for $(m,n)=(3,5)$) and Figure EC.3 (for $(m,n)=(6,10)$) in Appendix EC.5, and are denoted by $\mathcal{D}$, $\mathcal{C}^-$ and $\mathcal{C}$, respectively. $\mathcal{D}$ is analogous to the dedicated structure in the balanced system, $\mathcal{C}$ can be viewed as a generalized chaining structure created based on Lemma 3 (using $\mathcal{A}'=\mathcal{D}$) and $\mathcal{C}^-$ can be viewed as chaining minus one arc. We note that while $\mathcal{C}^-$ and $\mathcal{C}$ differ by just one arc, they differ significantly in GCG, leading us to anticipate a significant difference in average backlogs.

Similar to §6.1.1, to understand and compare the effectiveness of $\mathcal{C}^-$ and $\mathcal{C}$, we compute two performance measures, $R(\cdot)$ and $\Delta(\cdot)$, defined in (50). Recall that $R(\mathcal{C})$ represents the ratio between the backlog of $\mathcal{C}$ and that of full flexibility, and $\Delta(\mathcal{C})$ represents the ratio between the improvement (starting from $\mathcal{D}$) of $\mathcal{C}$ and that of full flexibility. The expected backlogs of $\mathcal{D}$, $\mathcal{C}$, $\mathcal{C}^-$ and full flexibility are evaluated using methods similar to those in §6.1.1. In Tables EC.4, EC.5 and EC.6, we present the values of $R(\mathcal{C})$, $\Delta(\mathcal{C})$, $B(\mathcal{C})$, $R(\mathcal{C}^-)$ and $\Delta(\mathcal{C}^-)$ for different system sizes.

The numerical results obtained in non-balanced and asymmetric systems suggest that the structure $\mathcal{C}$, which contains $m+n$ arcs and has a large GCG, is very effective. In general, the numerical results in asymmetric systems for $\mathcal{C}$ and $\mathcal{C}^-$ are similar to the numerical results in §6.1.1 for $\mathcal{LC}$ and $\mathcal{LC}^-$. For example, with ρ not close to 1, while $R(\mathcal{C})$ is often significantly greater than 1, $\mathcal{C}$ always performs strongly in the measure $\Delta(\cdot)$. More specifically, $\Delta(\mathcal{C})$ is at least 96.2% for all settings with $\rho=0.8$, and at least 92.3% over all of the tested settings. Also, similar to §6.1.1, the performance of $\mathcal{C}$ deteriorates as the system size increases. Thus, one should expect lower performance of $\mathcal{C}$ for systems with more than 20 plants and products.

Additionally, there is a significant difference between $\Delta(\mathcal{C})$ and $\Delta(\mathcal{C}^-)$ computationally, despite that $\mathcal{C}$ and $\mathcal{C}^-$ only differ by one arc. This confirms the insight we gained from our theoretical analysis in §4 and §5. That is, for an arbitrary unbalanced and asymmetric system, there exist systems where it is necessary to have $m+n$ arcs to create effective structures, and $m+n-1$ arcs are typically not enough.

6.2. Continuous-Time Environments

Past literature has observed the effectiveness of chaining in other dynamic environments such as production lines and call centers (see Hopp et al. (2004), Wallace and Whitt (2005), Iravani et al. (2005)). Motivated by these observations, we simulate the performances of the long chain in two different continuous-time environments and compare them with our findings in §6.1. The purpose of our simulation is not just to reproduce the results in the literature, but also to complement the previous simulation studies by varying parameters n and ρ. Also, in our simulation, we will

compute metrics $R(\cdot)$ and $\Delta(\cdot)$ for different structures, allowing us to directly compare numerical results from continuous-time settings with that from discrete-time.

6.2.1. Parallel Queueing Networks. We present the simulation results for long chain in the continuous-time parallel queueing environment, which is often used to simulate call centers (see, e.g., Wallace and Whitt (2005)). In the continuous-time parallel queueing environment, we have n types of customers that arrive continuously according to n mutually independent Poisson processes. The service time of each customer is distributed exponentially with rate 1, and there are a total of n servers. Similar to §6.1.1, we assume that the arrival rate is almost symmetric; type 1 customers have arrival rate 0.95ρ, type $2, 3, \ldots, n-1$ customers each have arrival rate ρ, and type n customers have arrival rate 1.05ρ. The parameters $n = 5, 10, 15$ and $\rho = 0.8, 0.9, 0.95, 0.975, 0.9875$. Finally, to simulate the performance of long chain, we adapt the discrete-time MW policy to the continuous environment, which, in our case, is equivalent to the longest-queue-first policy where each idling server serves the longest queue among the customer types it is capable of serving.

In Table EC.7, we present the performance of $\mathcal{LC}$ (the long chain) and $\mathcal{LC}^-$ (the long chain less arc $(n, 1)$) under different values of n and ρ. For each parameter setting, we simulate the system for 4500 warm-up time units, and then record the queue length for the next 500 time units, for 1000 randomly generated samples. (Note that the number of warm-up time units is much larger here compared to the discrete-time environment because in each time unit, we see on average a much smaller number of arrivals.) Same as in §6.1, $R(\mathcal{A})$ represents the ratio between the queue length of $\mathcal{A}$ and that of full flexibility; while $\Delta(\mathcal{A})$ represents the ratio between the improvement (starting from dedicated flexibility) of $\mathcal{A}$ and that of full flexibility. Finally, SE% denotes the maximum value of the standard error percentages among $\mathcal{D}$, $\mathcal{LC}$, $\mathcal{LC}^-$ and $\mathcal{F}$.

Table EC.7 shows that the performance of $\mathcal{LC}$ is significantly better than the performance of $\mathcal{LC}^-$ in all tested settings, echoing the observation made in Wallace and Whitt (2005) and Iravani et al. (2005) in parallel queueing networks. Similar to §6.1, with large n and ρ not close to 1, while the ratio between the queue length of $\mathcal{LC}$ to that of full flexibility is not always close to 1, the ratio $\Delta(\mathcal{A})$ between the improvement of $\mathcal{LC}$ to that of full flexibility is almost always better than 80%, indicating that going from the dedicated structure to $\mathcal{LC}$ provides most benefit. Also, because the performance of $\mathcal{LC}$ relative to full flexibility deteriorates as n increases, it implies that more flexibility than $\mathcal{LC}$ may be needed to further improve the system performance when n is large. This observation resonates with the theoretical findings of Tsitsiklis and Xu (2017), which shows that to achieve small backlogs, one needs a structure where the average degree for each node should scale as $\log n$ asymptotically when n is large. Finally, compared to the simulation results in §6.1, we see that the relative performance of $\mathcal{LC}$ in continuous parallel queueing networks is

worse compared to the discrete-time backlogging environments. Intuitively, this is because in the discrete-time system, arrivals and services can be thought of as being more "synchronized" in each time period, compared to those in the continuous-time setting, which makes the effectiveness of $\mathcal{LC}$ more pronounced.

6.2.2. Serial Production Line. Next, we present the simulation results for long chain in a continuous-time serial production line. We simulate an environment that was previously studied by Hopp et al. (2004), where flexible service stations operate under a constant work-in-process (CONWIP) release policy. Under this environment, a new job is released into the system only when a job is completed, hence keeping the total number of work-in-process at some fixed constant. Each new job requires n stages of processing before completion, where the processing time at each stage follows an exponential distribution. Similar to §6.1.1 and §6.2, we set the processing rate for each stage to be almost symmetric; stage 1 has a processing rate of 0.95, stage $2, 3, \ldots, n-1$ has a processing rate of 1, and stage n has a processing rate of 1.05. In addition, there are n service stations in the production line. A dedicated flexibility structure under this setting means that service station i is only capable of processing jobs at stage i, while the long chain structure has service station i capable of processing jobs at stage i and $i+1$ for $i = 1, ..., n-1$, and service station n is capable of processing jobs at stage n and stage 1. We vary $n = 5, 10, 15$ and the number of work-in-process (wip) $n, 2n, 5n$.

To simulate the performance of long chain, we use an adapted MW policy, where each service station process the longest backlogged stage that it is capable of serving. For each system, we first run 5000 time units as warm-up times for the system, and then record the number of jobs completed during the next 5000 time units, for 1000 randomly generated samples. Letting $P(\mathcal{A})$ denote the average number of jobs completed. We also compute two performance measures:

$$R(\mathcal{A}) = \frac{P(\mathcal{A})}{P(\mathcal{F})}, \quad \Delta(\mathcal{A}) = \frac{P(\mathcal{A}) - P(\mathcal{D})}{P(\mathcal{F}) - P(\mathcal{D})}, \quad \text{for any flexibility structure } \mathcal{A}. \tag{51}$$

where $R(\mathcal{A})$ represents the ratio between the number of jobs processed by $\mathcal{A}$ and that of full flexibility, and $\Delta(\mathcal{A})$ represents the ratio between the improvement (starting from dedicated flexibility) of $\mathcal{A}$ and that of full flexibility. Contrary to the previous settings, because the number of completed jobs in full flexibility is greater than that of $\mathcal{A}$, $R(\mathcal{A})$ is less than 1, while $\Delta(\mathcal{A})$ is greater than 1.

Table EC.8 presents the performance of $\mathcal{LC}$ (the long chain) and $\mathcal{LC}^-$ (the long chain less arc $(n, 1)$) under different values of n and the number of work-in-process (wip), measured by $R(\cdot)$ and $\Delta(\cdot)$. In the last column of Table EC.8, SE% denotes the maximum standard error percentages for the average number of jobs processed for dedicated, $\mathcal{LC}$, $\mathcal{LC}^-$ and full flexibility. There are some similarities between the performance of $\mathcal{LC}$ and $\mathcal{LC}^-$ in the serial production line to that of

the parallel queueing environments. In particular, $\mathcal{LC}$ significantly outperforms $\mathcal{LC}^-$ in all tested settings, and the effectiveness of $\mathcal{LC}$ relative to full flexibility seems to deteriorate as n grows. Also, the effectiveness of $\mathcal{LC}$ deteriorates when wip is low. This is intuitive because when wip is low, service stations under $\mathcal{LC}$ will tend to spend more time idling, while full flexibility will not have service stations idling as long as wip is larger than n.

7. Conclusion and Future Directions

To the best of our knowledge, this paper is the first to theoretically investigate the effectiveness of sparse flexibility structures in the multi-period systems with finite and unbalanced number of plants and products. We find that when capacity utilization is high, in order to achieve the similar performance as full flexibility in the multi-period MTO system, one only needs to design a sparse flexibility structure with $m+n$ arcs. Interestingly, we also find that all $m+n$ arcs are necessary to guarantee this type of asymptotic performance, as there exist systems that even the best structure with $m+n-1$ arcs cannot achieve the same asymptotic performance as full flexibility. In order to verify the robustness of our findings, we performed numerical experiments to understand the effectiveness chaining (and generalized chaining) structures. In short, the chaining structure performs very well when benchmarked against dedicated and full flexibility structures in multi-period make-to-order systems. However, the effectiveness of chaining deterioates as the system size grows.

To close our paper, we point out several interesting future research avenues. (a) *Non-uniform backlogging costs.* We have assumed in this paper that the per-unit backlogging cost is uniform across different products. The current methodology to establish our results does not readily extend to the case of non-uniform per-unit backlogging costs. We believe that techniques from, e.g., Ata and Kumar (2005), which considers non-uniform per-unit backlogging costs, may be combined with techniques from our paper to establish corresponding results for the case of non-uniform costs. (b) *Production system with inventories.* One can also study a production-inventory system where the firm uses the production resource to accumulate inventories in anticipation of future demand. We note that if one focus on the stationary base-stock policies suggested by Janakiraman et al. (2014), then one can view the difference between the inventory level and the base-stock level as "backlog", and potentially apply the tools we introduced in this paper. Two interesting open problems in this direction include, (i) the optimality of base-stock policies in the multi-product system with limited flexibility; and (ii) the effectiveness of sparse flexible systems under base-stock policies when the base-stock level is endogenous. (c) *Unknown demand rate.* The design of flexibility structures with sufficient GCG requires the knowledge of both the demand rate vector $\boldsymbol{\lambda}$ and the capacity vector $\boldsymbol{c}$. While it is reasonable to expect that most firms have full knowledge of $\boldsymbol{c}$, estimates of $\boldsymbol{\lambda}$ can

be uncertain or inaccurate. Therefore, it is worthwhile investigating the robustness of flexibility structure with respect to different input demand rates, and the trade-off between sparsity and inaccurate demand estimates. (d) *Finite-period models.* Our paper analyzed the infinite-horizon model. The finite-period flexible production models remain an open and important challenge. Our simulation suggests that the Max-weight policy with sufficient flexibility can work extremely well in finite horizon, and it would be very interesting to see any theoretical progress on this front.

Acknowledgments

The authors thank the area editor, the anonymous associate editor, and the anonymous referees for their constructive and detailed comments, which helped significantly improve both the content and the exposition of this paper. The research of Cong Shi is partially supported by a National Science Foundation (NSF) grant CMMI-1634505, and an MCubed grant at the University of Michigan at Ann Arbor.

References

Ahuja, R. K., T. L. Magnanti, J. B. Orlin. 1993. *Network flows*. Prentice Hall, Englewood Cliffs, NJ.

Andradóttir, S., H. Ayhan, D. G. Down. 2003. Dynamic server allocation for queueing networks with flexible servers. *Operations Research* **51**(6) 952–968.

Andradóttir, S., H. Ayhan, D. G. Down. 2007. Compensating for failures with flexible servers. *Operations Research* **55**(4) 753–768.

Andradóttir, S., H. Ayhan, D. G. Down. 2013. Design principles for flexible systems. *Production and Operations Management* **22**(5) 1144–1156.

Asadpour, A., X. Wang, J. Zhang. 2016. Online resource allocation with limited flexibility. Working Paper, New York University, NY.

Ata, B., S. Kumar. 2005. Heavy traffic analysis of open processing networks with complete resource pooling: Asymptotic optimality of discrete review policies. *The Annals of Applied Probability* **15**(1A) 331–391.

Bertsekas, D. P., S. E. Shreve. 2007. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, Cambridge, MA.

Bertsimas, D., D. Gamarnik, J. N. Tsitsiklis. 2001. Performance of multiclass markovian queueing networks via piecewise linear lyapunov functions. *Annals of Applied Probability* 1384–1428.

Cachon, G., C. Terwiesch. 2011. *Matching Supply with Demand: An Introduction to Operations Management*. 3rd ed. McGraw-Hill, Ashland, OH.

Chen, X., T. Ma, J. Zhang, Y. Zhou. 2016. Optimal design of process flexibility for general production systems. Working Paper, New York University, NY.

Chen, X., J. Zhang, Y. Zhou. 2015. Optimal sparse designs for process flexibility via probabilistic expanders. *Operations Research* **63**(5) 1159–1176.

Chou, M. C., G. A. Chua, C.-P. Teo, H. Zheng. 2010. Design for process flexibility: Efficiency of the long chain and sparse structure. *Operations Research* **58**(1) 43–58.

Chou, M. C., G. A. Chua, C.-P. Teo, H. Zheng. 2011. Process flexibility revisited: the graph expander and its applications. *Operations Research* **59**(5) 1090–1105.

Dai, J. G., W. Lin. 2005. Maximum pressure policies in stochastic processing networks. *Operations Research* **53**(2) 197–218.

Deng, T., Z.-J. M. Shen. 2013. Process flexibility design in unbalanced networks. *Manufacturing and Service Operations Management* **15**(1) 24–32.

Désir, A., V. Goyal, Y. Wei, J. Zhang. 2016. Sparse process flexibility designs: Is the long chain really optimal? *Operations Research* **64**(2) 416–431.

Eryilmaz, A., R. Srikant. 2012. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems* **72**(3-4) 311–359.

Gurvich, I., W. Whitt. 2009. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing & Service Operations Management* **11**(2) 237–253.

Hajek, B. 1982. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied probability* 502–525.

Harrison, J. M., M. J. López. 1999. Heavy traffic resource pooling in parallel-server systems. *Queueing systems* **33**(4) 339–368.

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

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

Janakiraman, G., M. Nagarajan, S. Veeraraghavan. 2014. Simple policies for managing flexible capacity. Working Paper, University of Texas at Dallas, TX.

Jordan, W., S. Graves. 1995. Principles on the benefits of manufacturing process flexibility. *Management Science* **41**(4) 577–594.

Keslassy, I., N. McKeown. 2001. Analysis of scheduling algorithms that provide 100% throughput in input-queued switches. *Proceedings of the Annual Allerton Conference on Communication Control and Computing*, vol. 39. Allerton House, Monticello, Illinois, 593–602.

Mandelbaum, A., A. L. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research* **52**(6) 836–855.

McKeown, N., A. Mekkittikul, V. Anantharam, J. Walrand. 1999. Achieving 100% throughput in an input-queued switch. *IEEE Transactions on Communications* **47**(8) 1260–1267.

Müller, A., D. Stoyan. 2002. *Comparison methods for stochastic models and risks*. John Wiley & Sons, Hoboken, New Jersey.

Powell, W. B. 2007. *Approximate Dynamic Programming: Solving the curses of dimensionality*. John Wiley & Sons, Hoboken, New Jersey.

Shah, D., D. Wischik. 2012. Switched networks with maximum weight policies: Fluid approximation and multiplicative state space collapse. *The Annals of Applied Probability* **22**(1) 70–127.

Sheng, L., H. Zheng, Y. Rong, W. T. Huh. 2015. Flexible system design: A perspective from service levels. *Operations Research Letters* **43**(3) 219–225.

Simchi-Levi, D. 2010. *Operations Rules: Delivering Customer Value through Flexible Operations*. MIT Press, Cambridge, MA.

Simchi-Levi, D., Y. Wei. 2012. Understanding the performance of the long chain and sparse designs in process flexibility. *Operations Research* **60**(5) 1125–1141.

Simchi-Levi, D., Y. Wei. 2015. Worst-case analysis of process flexibility designs. *Operations Research* **63**(1) 166–185.

Stolyar, A. L. 2004. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability* **14**(1) 1–53.

Tanrisever, F., D. Morrice, D. Morton. 2012. Managing capacity flexibility in make-to-order production environments. *European Journal of Operational Research* **216**(2) 334–345.

Tassiulas, L., A. Ephremides. 1992. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control* **37**(12) 1936–1948.

Tsitsiklis, J. N., K. Xu. 2017. Flexible queueing architectures. *Operations Research* **65**(5) 1398–1413.

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

Wang, X., J. Zhang. 2015. Process flexibility: A distribution-free bound on the performance of k-chain. *Operations Research* **63**(3) 555–571.

This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Electronic Companion to

"Process Flexibility for Multi-Period Production Systems"

by Shi, Wei and Zhong

EC.1. Stability Condition

Here, we characterize the conditions on the flexibility structure $\mathcal{A}$, the capacity vector $\boldsymbol{c}$ and the demand distribution $\mathbf{D}$ under which there exists a policy that guarantees the finiteness of the long-run average backlogging cost. The requirement of finite backlogging cost is also known as the *stability* condition, and this consideration motivates us to define the stability of a policy.

Definition EC.1 (Stable Policy). Given a flexibility structure $\mathcal{A}$, plant capacity $\mathbf{c}$, and product demand distribution $\mathbf{D}$, a production policy π is said to be *stable* if the expected long-run average backlogging costs $\Gamma(\pi) < \infty$, and is *unstable* otherwise.

The following proposition gives a necessary and sufficient condition on $\boldsymbol{c}$ and $\mathbf{D}$ for the existence of a stable policy under $\mathcal{A}$. The proof is standard, and is included here for completeness.

Proposition EC.1. *Let the flexibility structure $\mathcal{A}$ be given. Then, a necessary and sufficient condition for the existence of a stable policy is*

$$\sum_{\mathcal{S}_i \in N(\Omega)} c_i > \sum_{\mathcal{T}_j \in \Omega} \lambda_j, \textit{for all } \Omega \subseteq \{\mathcal{T}_1, \ldots, \mathcal{T}_n\},\ \Omega \neq \emptyset. \tag{EC.1}$$

Proof of Proposition EC.1. The necessity of (EC.1) can be derived as follows. For any non-empty subset $\Omega \subseteq \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$, the corresponding aggregate demand in time period t is $\sum_{\mathcal{T}_j \in \Omega} D_j(t)$, with $\mathbb{E}\left[\sum_{\mathcal{T}_j \in \Omega} D_j(t)\right] = \sum_{\mathcal{T}_j \in \Omega} \lambda_j$, and the maximum production capacity that can be devoted to this demand is $\sum_{\mathcal{S}_i \in N(\Omega)} c_i$. Consider a single-plant production system with capacity $\sum_{\mathcal{S}_i \in N(\Omega)} c_i$ and demand given by $\sum_{\mathcal{T}_j \in \Omega} D_j(t)$ in time period t, $t = 1, 2, \cdots$. Let $Q(t)$ be the backlog of the single-plant system at time t, and suppose that this single-plant system starts empty, i.e., $Q(0) = 0$. Then, by a standard coupling argument, it can be shown that $\sum_{\mathcal{T}_j \in \Omega} B_j(t) \geq Q(t)$ for each t. Thus, for the original system to be stable, we need $\limsup \frac{1}{T} \sum_{t=1}^{T} \sum_{\mathcal{T}_j \in \Omega} B_j(t) < \infty$, and so it is necessary that $\limsup \frac{1}{T} \sum_{t=1}^{T} Q(t) < \infty$. To guarantee $\limsup \frac{1}{T} \sum_{t=1}^{T} Q(t) < \infty$, it is required that $\sum_{\mathcal{S}_i \in N(\Omega)} c_i > \sum_{\mathcal{T}_j \in \Omega} \lambda_j$. This establishes necessity.

The sufficiency of (EC.1) can be proved as follows. Suppose that for each non-empty subset $\Omega \subseteq \{\mathcal{T}_1, \cdots, \mathcal{T}_n\}$, $\sum_{\mathcal{S}_i \in N(\Omega)} c_i > \sum_{\mathcal{T}_j \in \Omega} \lambda_j$. Scale the vector $\boldsymbol{\lambda}$ by a factor α so that for every non-empty subset $\tilde{\Omega} \subseteq \{\mathcal{T}_1, \cdots, \mathcal{T}_n\}$, $\sum_{\mathcal{S}_i \in N(\tilde{\Omega})} c_i \geq \alpha \sum_{\mathcal{T}_j \in \tilde{\Omega}} \lambda_j$, and there exists a non-empty subset $\Omega \subseteq \{\mathcal{T}_1, \cdots, \mathcal{T}_n\}$ such that $\sum_{\mathcal{S}_i \in N(\Omega)} c_i = \alpha \sum_{\mathcal{T}_j \in \Omega} \lambda_j$. Necessarily, $\alpha > 1$. By the max-flow min-cut

theorem, $\alpha\boldsymbol{\lambda} \in R(\mathcal{A})$. Let π be the production policy that uses the constant schedule $\alpha\boldsymbol{\lambda}$. Then, under the policy π, the system is stable. This concludes the sufficiency part. □

An immediate consequence of Proposition EC.1 is the following corollary.

COROLLARY EC.1. *A necessary and sufficient condition for the existence of a stable policy under the full flexibility structure is* $\Lambda < C$.

Corollary EC.1 essentially spells out that the weakest condition for the existence of a stable policy under any given flexibility structure is $\Lambda < C$, justifying Equation (3) in Assumption 1.

EC.2. Proofs of Results in §3.2

EC.2.1. Proof of Proposition 1

A key fact that is used in the proof of Proposition 1 is the following lemma.

LEMMA EC.1. *Let the demand rate vector* $\boldsymbol{\lambda}$ *and capacity vector* $\mathbf{c}$ *be given with* $\Lambda < C$, *and* $\boldsymbol{\lambda}'$ *be the projection of* $\boldsymbol{\lambda}$ *onto the plane defined by* $\left\{\mathbf{g} \mid \sum_{j=1}^{n} g_j = C\right\}$. *Let* $\mathcal{A}$ *be a flexibility structure that has positive GCG* ($\eta > 0$). *Then, for any* $\mathbf{x} \in \mathbb{R}^n$ *with* $\sum_j x_j = 0$ *and* $\|\mathbf{x}\| \le \eta/\sqrt{n}$, $\boldsymbol{\lambda}' + \mathbf{x}$ *lies on the face defined by* $\left\{\mathbf{g} \mid \sum_{j=1}^{n} g_j = C\right\}$, *and* $\boldsymbol{\lambda}' + \mathbf{x} \in R(\mathcal{A})$.

Proof of Lemma EC.1. Let $\mathbf{x} \in \mathbb{R}^n$ be such that $\sum_j x_j = 0$ and $\|\mathbf{x}\| \le \eta/\sqrt{n}$. Then, $\sum_j |x_j| \le \eta$ by the Cauchy-Schwarz inequality. Second, since $\sum_j x_j = 0$, $\sum_j (\lambda'_j + x_j) = \sum_j \lambda'_j = C$, and $\boldsymbol{\lambda}' + \mathbf{x}$ lies on the face defined by $\left\{\mathbf{g} \mid \sum_{j=1}^{n} g_j = C\right\}$. Finally, for any $\Omega \subsetneq \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$,

$$\sum_{\mathcal{T}_j \in \Omega} (\lambda'_j + x_j) = \sum_{\mathcal{T}_j \in \Omega} \lambda'_j + \sum_{\mathcal{T}_j \in \Omega} x_j \le \sum_{\mathcal{T}_j \in \Omega} \lambda'_j + \sum_j |x_j| \le \left(\sum_{\mathcal{S}_i \in N(\Omega)} c_i - \eta \right) + \eta = \sum_{\mathcal{S}_i \in N(\Omega)} c_i.$$

We also note that for each j, $\lambda'_j + x_j \ge 0$. To see this, for each j, let $\Omega^{-j} = \{\mathcal{T}_1, \ldots \mathcal{T}_{j-1}, \mathcal{T}_{j+1}, \ldots, \mathcal{T}_n\}$, and we have

$$\eta \le \sum_{\mathcal{S}_i \in N(\Omega^{-j})} c_i - \sum_{\mathcal{T}_j \in \Omega^{-j}} \lambda'_j \le \sum_{i=1}^{m} c_i - \sum_{\mathcal{T}_j \in \Omega^{-j}} \lambda'_j = \lambda'_j.$$

Thus, we have that $\lambda'_j \ge \eta \ge |x_j|$, implying $\lambda'_j + x_j \ge 0$. As a result, we can conclude that $\boldsymbol{\lambda}' + \mathbf{x} \in R(\mathcal{A})$. □

Lemma EC.1 states if $\mathcal{A}$ has a positive GCG η, the Euclidean ball defined on the hyperplane $\left\{\mathbf{g} \mid \sum_{j=1}^{n} g_j = C\right\}$ with center $\boldsymbol{\lambda}'$ and radius $\eta/\sqrt{n}$, lies within the production polytope $R(\mathcal{A})$. The lemma therefore allows us to connect GCG to the result in Eryilmaz and Srikant (2012), which is used for the proof of Proposition 1.

We first note that the system is stable under the Max-Weight policy. The proof of this fact is quite standard, by considering the conditional expected drift of the quadratic Lyapunov function $\sum_{j=1}^{n} B_j^2$, and invoking the so-called Foster's lemma. Similar proofs have appeared in e.g., McKeown et al. (1999), Tassiulas and Ephremides (1992), Dai and Lin (2005). We skip details.

Since the Max-Weight policy is stable, there exists a unique steady-state distribution. Let $\mathbf{B}(\infty)$ be the unique random backlog vector in steady state. Furthermore, for any backlog vector $\mathbf{B} = (B_1, B_2, \cdots, B_n)$, define $\bar{B} = (B_1 + B_2 + \cdots + B_n)/n$ to be the average of the backlogs, and let $\mathbf{\Delta B} = \mathbf{B} - \bar{B}\mathbf{1}$, where $\mathbf{1} = (1, 1, \ldots, 1)$. Note that $\mathbf{\Delta B}$ is the vector of deviations of the backlogs from their average. In addition, we define $D_{\max}$ as the maximum possible value for the aggregate demand. Under Assumption 1, we have $D_{\max} \le nu$.

To prove Proposition 1, we first establish the following state space collapse result. Informally, state space collapse implies that under the Max-Weight policy, all backlogs B_i stays close to the mean $\bar{B}$, so that the vector $\mathbf{\Delta B}$ remains small.

THEOREM EC.1 **(State space collapse)**. *Let $\mathbf{B}(\infty)$ have the steady-state distribution of the backlog vector under Max-Weight policy, and flexibility structure $\mathcal{A}$ with the GCG η. Then, for any $\ell \in \mathbb{Z}_+$,*

$$\mathbb{P}\left(\|\mathbf{\Delta B}(\infty)\|_2 > K' + 2\xi\ell\right) \le \left(\frac{\xi}{\xi+\gamma}\right)^{\ell+1}, \tag{EC.2}$$

where

$$K' = \frac{\left(\Sigma^2 + \sum_j \lambda_j^2 + 2C^2\right)\sqrt{n}}{\eta}; \quad \gamma = \frac{\eta}{2\sqrt{n}}; \quad \xi = \sqrt{n}(D_{\max} + C). \tag{EC.3}$$

Proof of Theorem EC.1. The proof of Theorem EC.1 invokes the following theorem in Bertsimas et al. (2001) (where a weaker version was also given in Hajek (1982)).

THEOREM EC.2. *Let $X(\cdot)$ be an irreducible, aperiodic and positively recurrent discrete-time Markov chain with a countable state space $\mathcal{X}$. Suppose that there exists a Lyapunov function $\Phi : \mathcal{X} \to \mathbb{R}_+$ with the following properties.*

(a) **Bounded increment.** *There exists a positive constant ξ such that $|\Phi(X(t+1)) - \Phi(X(t))| \le \xi$ for all t a.s.*

(b) **Negative drift.** *There exist positive constants K' and γ such that whenever $\Phi(X(t)) > K'$,*

$$\mathbb{E}\left[\Phi(X(t+1)) - \Phi(X(t)) \mid X(t)\right] \le -\gamma. \tag{EC.4}$$

Then, under the steady-state distribution of $X(\cdot)$, for any $\ell \in \mathbb{Z}_+$,

$$\mathbb{P}\left(\Phi(X) > K' + 2\xi\ell\right) \le \left(\frac{\xi}{\xi+\gamma}\right)^{\ell+1}. \tag{EC.5}$$

The proof of Theorem EC.1 then relies on establishing conditions (a) and (b) of Theorem EC.2 for an appropriately chosen Lyapunov function $\Phi(\cdot)$. It is straightforward to check that $\mathbf{B}(\cdot)$ is aperiodic and positively recurrent. We also claim that if the product demand instances and plant capacities are integral (rational), then the state space of $\mathbf{B}(\cdot)$ is countable. To see this, note that the Max-Weight policy solves a network flow problem with integral (rational) input at each time period, implying that if $\mathbf{b}(t-1)$ and $\mathbf{d}(t)$ are integral (rational), then $\mathbf{b}(t)$ is integral (rational). Instead of focusing on the Markov chain $\mathbf{B}(\cdot)$, we will consider the closely related chain $\mathbf{B}'(\cdot)$, defined to be $\mathbf{B}'(t) = \mathbf{B}(t) + \mathbf{D}(t)$ for all t.

PROPOSITION EC.2. *The following inequality holds for the Lyapunov function* $\Phi(\mathbf{B}') = \|\Delta\mathbf{B}'\|_2$.

$$\mathbb{E}\left[\|\Delta\mathbf{B}'(t+1)\|_2 - \|\Delta\mathbf{B}'(t)\|_2 \mid \mathbf{B}'(t)\right] \le -\frac{\eta}{\sqrt{n}} + \frac{\Sigma^2 + \sum_j \lambda_j^2 + 2C^2}{2\|\Delta\mathbf{B}'(t)\|_2}. \tag{EC.6}$$

In particular, whenever $\|\Delta\mathbf{B}'(t)\|_2 > \frac{\sqrt{n}\left(\Sigma^2+\sum_j \lambda_j^2+2C^2\right)}{\eta}$,

$$\mathbb{E}\left[\|\Delta\mathbf{B}'(t+1)\|_2 - \|\Delta\mathbf{B}'(t)\|_2 \mid \mathbf{B}'(t)\right] \le -\frac{\eta}{2\sqrt{n}}. \tag{EC.7}$$

Proof of Proposition EC.2. The proof of Proposition EC.2 mainly consists of establishing the following expressions.

(a) Show that

$$\|\Delta\mathbf{B}'(t+1)\|_2 - \|\Delta\mathbf{B}'(t)\|_2 \le \frac{\left(\|\mathbf{B}'(t+1)\|_2^2 - \|\mathbf{B}'(t)\|_2^2\right) - n\left(\overline{B'(t+1)}^2 - \overline{B'(t)}^2\right)}{2\|\Delta\mathbf{B}'(t)\|_2}; \tag{EC.8}$$

(b) show that

$$\mathbb{E}\left[\|\mathbf{B}'(t+1)\|_2^2 - \|\mathbf{B}'(t)\|_2^2 \mid \mathbf{B}'(t)\right] \le -2n\zeta\overline{B'(t)} - \frac{2\eta}{\sqrt{n}}\|\Delta\mathbf{B}'(t)\|_2 + \left(\Sigma^2 + C^2 + \sum_j \lambda_j^2\right); \tag{EC.9}$$

and

(c) show that

$$\mathbb{E}\left[n\left(\overline{B'(t+1)}^2 - \overline{B'(t)}^2\right) \mid \mathbf{B}'(t)\right] \ge -2n\zeta\overline{B'(t)} + \frac{1}{n}\left(\Sigma^2 + n^2\zeta^2 - 3C^2\right). \tag{EC.10}$$

(a) To establish (EC.8), we use the following general inequality: if $x > 0$, then $y - x \le \frac{y^2-x^2}{2x}$. Substituting $\|\Delta\mathbf{B}'(t)\|_2$ in place of x and $\|\Delta\mathbf{B}'(t+1)\|_2$ in place of y, we get

$$\|\Delta\mathbf{B}'(t+1)\|_2 - \|\Delta\mathbf{B}'(t)\|_2 \le \frac{\|\Delta\mathbf{B}'(t+1)\|_2^2 - \|\Delta\mathbf{B}'(t)\|_2^2}{2\|\Delta\mathbf{B}'(t)\|_2}. \tag{EC.11}$$

It is easy to verify that in general, $\langle \Delta\mathbf{B}, \bar{B}\mathbf{1}\rangle = 0$ and $\Delta\mathbf{B} + \bar{B}\mathbf{1} = \mathbf{B}$. Thus, by the Pythagorean theorem,

$$\|\Delta\mathbf{B}'(t)\|_2^2 = \|\mathbf{B}'(t)\|_2^2 - \|\overline{B'(t)}\mathbf{1}\|_2^2 = \|\mathbf{B}'(t)\|_2^2 - n\overline{B'(t)}^2.$$

A similar identity holds for $\|\Delta\mathbf{B}'(t+1)\|_2^2$. Therefore, substituting these identities into (EC.11), we establish (EC.8).

(b) To establish (EC.9), we first write $\mathbf{B}'(t+1) = \mathbf{B}'(t) - \mathbf{G}(\mathbf{B}'(t)) + \mathbf{U}(\mathbf{B}'(t)) + \mathbf{D}(t+1)$. Here $\mathbf{G}(\mathbf{B}'(t))$ is the production schedule used at time t, which depends on the vector $\mathbf{B}'(t)$. We use the following convention here: we suppose that $\mathbf{G}(\mathbf{B}'(t))$ is obtained from some $(f_{ij})_{i,j}$ such that $\sum_j f_{ij} = c_i$ for all i; i.e., all plants use their production capability fully in each time period. It may happen that for some j, $B_j'(t) < G_j(\mathbf{B}'(t))$, in which case we let $U_j(t) = G_j(\mathbf{B}'(t)) - B_j'(t)$ be the unused capacity for product j. Otherwise, let $U_j(t) = 0$. Then, we denote the vector $(U_j(t))_j$ by $\mathbf{U}(\mathbf{B}'(t))$. An immediate consequence is that $\langle \mathbf{B}(t+1), \mathbf{U}(t)\rangle = \langle \mathbf{B}'(t) - \mathbf{G}(\mathbf{B}'(t)) + \mathbf{U}(t), \mathbf{U}(t)\rangle = 0$.

For notational convenience, we drop the time index, and then

$$\mathbb{E}\left[\|\mathbf{B}'(t+1)\|_2^2 - \|\mathbf{B}'(t)\|_2^2 \,|\, \mathbf{B}'(t)\right] = \mathbb{E}\left[\|\mathbf{B}' - \mathbf{G} + \mathbf{U} + \mathbf{D}\|_2^2 - \|\mathbf{B}'\|_2^2 \,|\, \mathbf{B}'\right].$$

We now focus on the term $\mathbb{E}[\|\mathbf{B}' - \mathbf{G} + \mathbf{U} + \mathbf{D}\|_2^2 - \|\mathbf{B}'\|_2^2 \,|\, \mathbf{B}']$. We have

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{B}' - \mathbf{G} + \mathbf{U} + \mathbf{D}\|_2^2 - \|\mathbf{B}'\|_2^2 \,|\, \mathbf{B}'\right] \\
&= \mathbb{E}\left[\|\mathbf{B}' - \mathbf{G} + \mathbf{D}\|_2^2 \,|\, \mathbf{B}'\right] + \mathbb{E}\left[\|\mathbf{U}\|_2^2 \,|\, \mathbf{B}'\right] + 2\mathbb{E}\left[\langle \mathbf{D}, \mathbf{U}\rangle \,|\, \mathbf{B}'\right] - \mathbb{E}\left[2\|\mathbf{U}\|_2^2 \,|\, \mathbf{B}'\right] - \mathbb{E}\left[\|\mathbf{B}'\|_2^2 \,|\, \mathbf{B}'\right] \\
&= \mathbb{E}\left[2\langle \mathbf{D} - \mathbf{G}, \mathbf{B}'\rangle \,|\, \mathbf{B}'\right] + \mathbb{E}\left[\|\mathbf{D} - \mathbf{G}\|_2^2 \,|\, \mathbf{B}'\right] - \mathbb{E}\left[\|\mathbf{U}\|_2^2 \,|\, \mathbf{B}'\right] + 2\mathbb{E}\left[\langle \mathbf{D}, \mathbf{U}\rangle \,|\, \mathbf{B}'\right] \\
&\le \mathbb{E}\left[2\langle \mathbf{D} - \mathbf{G}, \mathbf{B}'\rangle \,|\, \mathbf{B}'\right] + \mathbb{E}\left[\|\mathbf{D} - \mathbf{G}\|_2^2 \,|\, \mathbf{B}'\right] + 2\mathbb{E}\left[\langle \mathbf{D}, \mathbf{U}\rangle \,|\, \mathbf{B}'\right], \qquad \text{(EC.12)}
\end{aligned}$$

where the first equality holds because $\mathbb{E}[\langle \mathbf{B}' - \mathbf{G}, \mathbf{U}\rangle \,|\, \mathbf{B}'] = -\mathbb{E}[\|\mathbf{U}\|_2^2 \,|\, \mathbf{B}']$. Next, let us start with $\mathbb{E}[\langle \mathbf{D} - \mathbf{G}, \mathbf{B}'\rangle \,|\, \mathbf{B}']$ and observe that

$$\begin{aligned}
\mathbb{E}[\langle \mathbf{D} - \mathbf{G}, \mathbf{B}'\rangle \,|\, \mathbf{B}'] &= \langle \boldsymbol{\lambda} - \mathbf{G}, \mathbf{B}'\rangle \\
&= \langle \boldsymbol{\lambda}' - \mathbf{G}, \mathbf{B}'\rangle - \langle \boldsymbol{\lambda}' - \boldsymbol{\lambda}, \mathbf{B}'\rangle \\
&= \langle \boldsymbol{\lambda}' - \mathbf{G}, \mathbf{B}'\rangle - n\zeta\overline{B'}.
\end{aligned}$$

By the Max-Weight policy, $\mathbf{G}$ is chosen from $R(\mathcal{A})$ to maximize the inner product $\langle \mathbf{G}, \mathbf{B}'\rangle$. By Lemma EC.1, $\boldsymbol{\lambda}' + \frac{\eta}{\sqrt{n}} \cdot \frac{\Delta\mathbf{B}'}{\|\Delta\mathbf{B}'\|} \in R(\mathcal{A})$. Therefore, we have

$$\begin{aligned}
\mathbb{E}[\langle \mathbf{D} - \mathbf{G}, \mathbf{B}'\rangle \,|\, \mathbf{B}'] &= \langle \boldsymbol{\lambda}' - \mathbf{G}, \mathbf{B}'\rangle - n\zeta\overline{B'} \\
&\le -\frac{\eta}{\sqrt{n} \cdot \|\Delta\mathbf{B}'\|}\langle \Delta\mathbf{B}', \mathbf{B}'\rangle - n\zeta\overline{B'}
\end{aligned}$$

$$= -\frac{\eta}{\sqrt{n}\cdot\|\Delta\mathbf{B}'\|}\langle\Delta\mathbf{B}',\Delta\mathbf{B}'\rangle - n\zeta\overline{B'}$$
$$= -\frac{\eta}{\sqrt{n}}\|\Delta\mathbf{B}'\| - n\zeta\overline{B'}. \tag{EC.13}$$

For the remaining terms, we have

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{D}-\mathbf{G}\|_2^2 \mid \mathbf{B}'\right] + 2\mathbb{E}\left[\langle\mathbf{D},\mathbf{U}\rangle \mid \mathbf{B}'\right] &= \mathbb{E}\left[\sum_j (D_j-G_j)^2 \mid \mathbf{B}'\right] + 2\mathbb{E}\left[\sum_j D_jU_j \mid \mathbf{B}'\right]\\ &= \mathbb{E}\left[\sum_j D_j^2 + \sum_j G_j^2 \mid \mathbf{B}'\right] - 2\mathbb{E}\left[\sum_j D_jG_j \mid \mathbf{B}'\right] + 2\mathbb{E}\left[\sum_j D_jU_j \mid \mathbf{B}'\right]\\ &\le \mathbb{E}\left[\sum_j D_j^2 + \sum_j G_j^2 \mid \mathbf{B}'\right] \le \Sigma^2 + \sum_j \lambda_j^2 + C^2. \end{aligned}\tag{EC.14}$$

Combining (EC.12), (EC.13) and (EC.14), we have established (EC.9).

(c) To establish (EC.10), we have

$$\begin{aligned}&\mathbb{E}\left[n\left(\overline{B'(t+1)}^2 - \overline{B'(t)}^2\right) \mid \mathbf{B}'(t)\right]\\ &= \mathbb{E}\left[\frac{1}{n}\left(\sum_j B_j'(t+1)\right)^2 - \frac{1}{n}\left(\sum_j B_j'(t)\right)^2 \,\Big|\, \mathbf{B}'(t)\right]\\ &= \frac{1}{n}\mathbb{E}\left[\left(\sum_j (B_j' - G_j + U_j + D_j)\right)^2 - \left(\sum_j B_j'\right)^2 \,\Big|\, \mathbf{B}'\right]\\ &= \frac{2}{n}\left(\sum_j B_j'\right)\mathbb{E}\left[\sum_j (D_j-G_j)\Big|\, \mathbf{B}'\right] + \frac{1}{n}\mathbb{E}\left[\left(\sum_j (D_j-G_j)\right)^2 \,\Big|\, \mathbf{B}'\right]\\ &\quad + \frac{2}{n}\mathbb{E}\left[\left(\sum_j (B_j'+D_j-G_j)\right)\left(\sum_j U_j\right) \,\Big|\, \mathbf{B}'\right]\\ &= -2n\zeta\overline{B'} + \frac{1}{n}\left(\Sigma^2 + (\Lambda-C)^2\right) + \frac{2}{n}\mathbb{E}\left[\left(\sum_j (B_j'+D_j)\right)\left(\sum_j U_j\right) \,\Big|\, \mathbf{B}'\right] - \frac{2C}{n}\mathbb{E}\left[\sum_j U_j\Big|\, \mathbf{B}'\right]\\ &\ge -2n\zeta\overline{B'} + \frac{1}{n}\left(\Sigma^2 + (\Lambda-C)^2\right) - \frac{2C^2}{n} = -2n\zeta\overline{B'} + \frac{1}{n}\left(\Sigma^2 + n^2\zeta^2 - 3C^2\right).\end{aligned}$$

Combining (EC.9) and (EC.10), we have

$$\begin{aligned}&\mathbb{E}\left[\left(\|\mathbf{B}'(t+1)\|_2^2 - \|\mathbf{B}'(t)\|_2^2\right) - n\left(\overline{B'(t+1)}^2 - \overline{B'(t)}^2\right) \,\Big|\, \mathbf{B}'(t)\right]\\ &\le -2n\zeta\overline{B'(t)} - \frac{2\eta}{\sqrt{n}}\|\Delta\mathbf{B}'(t)\|_2 + \left(\Sigma^2 + C^2 + \sum_j \lambda_j^2\right) + 2n\zeta\overline{B'(t)} - \frac{1}{n}\left(\Sigma^2 + n^2\zeta^2 - 3C^2\right)\\ &\le -\frac{2\eta}{\sqrt{n}}\|\Delta\mathbf{B}'(t)\|_2 + \left(\Sigma^2 + 2C^2 + \sum_j \lambda_j^2\right).\end{aligned}$$

Thus, by (EC.8),

$$\mathbb{E}\left[\|\Delta\mathbf{B}'(t+1)\|_2 - \|\Delta\mathbf{B}'(t)\|_2 \,|\, \mathbf{B}'(t)\right] \le \frac{1}{2\|\Delta\mathbf{B}'(t)\|_2}\left(-\frac{2\eta}{\sqrt{n}}\|\Delta\mathbf{B}'(t)\|_2 + \left(\Sigma^2 + 2C^2 + \sum_j \lambda_j^2\right)\right)$$
$$\le -\frac{\eta}{\sqrt{n}} + \frac{\Sigma^2 + \sum_j \lambda_j^2 + 2C^2}{2\|\Delta\mathbf{B}'(t)\|_2}.$$

This concludes the proof of Proposition EC.2. □

With Proposition EC.2, we can now complete the proof of Theorem EC.1.

Proof of Theorem EC.1. We have already established the negative drift condition (b) (of Theorem EC.2) in Proposition EC.2. To establish condition (a), first note that

$$\left|\|\Delta\mathbf{B}'(t+1)\|_2 - \|\Delta\mathbf{B}'(t)\|_2\right| \le \left|\|\mathbf{B}'(t+1)\|_2 - \|\mathbf{B}'(t)\|_2\right|.$$

Second, the maximum decrease in each $B_j'(t)$ is C, and the maximum increase in each $B_j'(t)$ is $D_{\max}$. Therefore, almost surely, for each j,

$$|B_j'(t+1) - B_j'(t)| \le C + D_{\max}.$$

This implies that almost surely, for every t,

$$\left|\|\mathbf{B}'(t+1)\|_2 - \|\mathbf{B}'(t)\|_2\right| \le \|(C + D_{\max})\mathbf{1}\|_2 = \sqrt{n}(C + D_{\max}).$$

By setting $K' = \sqrt{n}\left(\sum_j \lambda_j^2 + 2C^2\right)/\eta$, $\gamma = \eta/(2\sqrt{n})$, and $\xi = \sqrt{n}(C + D_{\max})$, and invoking Theorem EC.2, we can establish Theorem EC.1. □

An immediate consequence of Theorem EC.1 is that all moments of $\|\Delta\mathbf{B}(\infty)\|_2$ are finite, and that $\mathbb{E}[\|\Delta\mathbf{B}(\infty)\|_2^2]$ is of order $O(1/\eta^2)$. There is also an immediate corollary to the state space collapse result in Theorem EC.1.

Corollary EC.2. *Let the setup be the same as in Theorem EC.1. Then,*

$$\mathbb{E}[\|\Delta\mathbf{B}\|_2^2] \le \left(\frac{\sqrt{n}\left(\Sigma^2 + \sum_j \lambda_j^2\right) + 14n^{3/2}C^2 + 12n^{3/2}D_{\max}^2}{\eta} + \sqrt{n}(D_{\max} + C)\right)^2. \tag{EC.15}$$

We now proceed to the formal proof of Proposition 1.

Proof of Proposition 1. Similar to the proof of Theorem EC.1, let $\mathbf{B}(\infty)$ be a random vector that has the stationary distribution of the Markov chain $\mathbf{B}(\cdot)$ under the Max-Weight policy. We are interested in the steady-state expected total backlog $\mathbb{E}\left[\sum_j B_j(\infty)\right]$. Suppose that at time -1, the initial backlog vector has the distribution of $\mathbf{B}(\infty)$. We will focus instead on the in-period backlog

vector $\mathbf{B}'(\infty) \triangleq \mathbf{B}(\infty) + \mathbf{D}(0)$, where $\mathbf{D}(0)$ is the random demand vector in period 0, realized after $\mathbf{B}(\infty)$. Now, consider the backlog vector $\mathbf{B}^+(\infty)$, and the in-period backlog vector $\mathbf{B}'^+(\infty)$, both in time period 1. Then, by stationarity, $\mathbf{B}^+(\infty)$ and $\mathbf{B}'(\infty)$ have the same distribution, and we can write

$$\mathbf{B}^+(\infty) = \mathbf{B}'(\infty) - \mathbf{G}(\mathbf{B}'(\infty)) + \mathbf{U}(\mathbf{B}'(\infty)), \quad \text{and} \quad \mathbf{B}'^+(\infty) = \mathbf{B}^+(\infty) + \mathbf{D}(1),$$

where $\mathbf{G}(\mathbf{B}'(\infty))$ is the production allocation under the Max-Weight policy, based on the updated backlog vector $\mathbf{B}'(\infty)$, $\mathbf{U}(\mathbf{B}'(\infty))$ is the vector of unused capacities, and $\mathbf{D}(1)$ the random demand vector in time period 1, realized immediately after $\mathbf{B}^+(\infty)$. With a slight abuse of notation, and to simplify notation, we write $\mathbf{D}$ for $\mathbf{D}(1)$ for the rest of this section. It is useful to note that because $\mathbf{D}$ is the demand vector in time period 1, but not time period 0, $\mathbf{D}$ is independent from $\mathbf{B}'(\infty)$, $\mathbf{B}^+(\infty)$, and $\mathbf{U}(\mathbf{B}'(\infty))$.

Write $\mathbf{G}(\infty) = \mathbf{G}(\mathbf{B}'(\infty))$ and $\mathbf{U}(\infty) = \mathbf{U}(\mathbf{B}'(\infty))$. Then, by stationarity,

$$\mathbb{E}\left[\left(\sum_j B_j'^+(\infty)\right)^2\right] = \mathbb{E}\left[\left(\sum_j B_j'(\infty)\right)^2\right],$$

and

$$\begin{aligned}
0 &= \mathbb{E}\left[\left(\sum_j B_j'^+(\infty)\right)^2 - \left(\sum_j B_j'(\infty)\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_j B'(\infty) - \sum_j G_j(\infty) + \sum_j U_j(\infty) + \sum_j D_j\right)^2 - \left(\sum_j B_j'(\infty)\right)^2\right] \\
&= 2\mathbb{E}\left[\left(\sum_j D_j - \sum_j G_j(\infty)\right)\left(\sum_j B_j'(\infty)\right)\right] + \mathbb{E}\left[\left(\sum_j D_j - \sum_j G_j(\infty)\right)^2\right] \\
&\quad + \mathbb{E}\left[\left(\sum_j U_j(\infty)\right)^2\right] + 2\mathbb{E}\left[\left(\sum_j B_j'(\infty) - \sum_j G_j(\infty) + \sum_j D_j\right)\left(\sum_j U_j(\infty)\right)\right] \\
&= 2\mathbb{E}\left[\left(\sum_j D_j - \sum_j G_j(\infty)\right)\left(\sum_j B_j'(\infty)\right)\right] + \mathbb{E}\left[\left(\sum_j D_j - \sum_j G_j(\infty)\right)^2\right] \\
&\quad + \mathbb{E}\left[\left(\sum_j U_j(\infty)\right)^2\right] + 2\mathbb{E}\left[\left(\sum_j B_j'^+(\infty) - \sum_j U_j(\infty)\right)\left(\sum_j U_j(\infty)\right)\right] \\
&= 2\mathbb{E}\left[\left(\sum_j D_j - \sum_j G_j(\infty)\right)\left(\sum_j B_j'(\infty)\right)\right] + \mathbb{E}\left[\left(\sum_j D_j - \sum_j G_j(\infty)\right)^2\right]
\end{aligned}$$

$$+2\mathbb{E}\left[\left(\sum_j B_j'^{+}(\infty)\right)\left(\sum_j U_j(\infty)\right)\right]-\mathbb{E}\left[\left(\sum_j U_j(\infty)\right)^2\right].$$

Write $\tilde{D}=\sum_j D_j$, and note that $\sum_j G_j(\infty)=C$. Then,

$$\begin{aligned}\mathbb{E}\left[\left(\sum_j D_j-\sum_j G_j(\infty)\right)\left(\sum_j B_j'(\infty)\right)\right] &= \mathbb{E}\left[(\tilde{D}-C)\sum_j B_j'(\infty)\right]\\ &= (\Lambda-C)\mathbb{E}\left[\sum_j B_j'(\infty)\right]=-n\zeta\mathbb{E}\left[\sum_j B_j'(\infty)\right],\end{aligned}$$

where the second inequality follows from the independence between $\mathbf{D}$ and $\mathbf{B}'(\infty)$. Therefore,

$$\begin{aligned}2n\zeta\mathbb{E}\left[\sum_j B_j'(\infty)\right] &= \mathbb{E}\left[\left(\tilde{D}-C\right)^2\right]+2\mathbb{E}\left[\left(\sum_j B_j'^{+}(\infty)\right)\left(\sum_j U_j(\infty)\right)\right]-\mathbb{E}\left[\left(\sum_j U_j(\infty)\right)^2\right]\\ &\le \mathbb{E}\left[\left(\tilde{D}-C\right)^2\right]+2\mathbb{E}\left[\left(\sum_j B_j'^{+}(\infty)\right)\left(\sum_j U_j(\infty)\right)\right]-\mathbb{E}\left[\left(\sum_j U_j(\infty)\right)\right]^2\\ &\le \mathbb{E}\left[\left(\tilde{D}-C\right)^2\right]+2\mathbb{E}\left[\left(\sum_j B_j'^{+}(\infty)\right)\left(\sum_j U_j(\infty)\right)\right]-n^2\zeta^2. \qquad \text{(EC.16)}\end{aligned}$$

We now analyze the first two terms on the right-hand side separately.

(a) $\mathbb{E}\left[\left(\tilde{D}-C\right)^2\right]$. Noting that $\mathbb{E}[\tilde{D}]=\Lambda$ and $Var[\tilde{D}]=\sum_j\sigma_j^2=\Sigma^2$, we have

$$\mathbb{E}\left[\left(\tilde{D}-C\right)^2\right]=\Sigma^2+(C-\Lambda)^2=\Sigma^2+n^2\zeta^2. \qquad \text{(EC.17)}$$

(b) $\mathbb{E}\left[\left(\sum_j B_j'^{+}(\infty)\right)\left(\sum_j U_j(\infty)\right)\right]$. First, we have

$$\mathbb{E}\left[\left(\sum_j B_j'^{+}(\infty)\right)\left(\sum_j U_j(\infty)\right)\right]=\mathbb{E}\left[\left(\sum_j B_j^{+}(\infty)+\tilde{D}\right)\left(\sum_j U_j(\infty)\right)\right].$$

Second, $\tilde{D}$ and $\mathbf{U}(\infty)$ are independent, so

$$\mathbb{E}\left[\tilde{D}\sum_j U_j(\infty)\right]=\Lambda\mathbb{E}\left[\sum_j U_j(\infty)\right]=\Lambda(C-\Lambda)=\Lambda n\zeta.$$

We now consider the term $\mathbb{E}\left[\left(\sum_j B_j^{+}(\infty)\right)\left(\sum_j U_j(\infty)\right)\right]$. Let us first note that for each j,

$B_j^{+}(\infty)U_j(\infty)=0$, since if there were any unused capacity for product j (i.e., $U_j(\infty)>0$), there

would be no backlog after production (i.e., $B_j^+(\infty)=0$). Thus,

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left[\left(\sum_j B_j^+(\infty)\right)\left(\sum_j U_j(\infty)\right)\right] &= \mathbb{E}\left[\overline{B^+(\infty)}\sum_j U_j(\infty)\right]\\ &= \mathbb{E}\left[\sum_j \overline{B^+(\infty)}U_j(\infty) - \sum_j B_j^+(\infty)U_j(\infty)\right]\\ &= \mathbb{E}\left[\sum_j (\overline{B^+(\infty)} - B_j^+(\infty))U_j(\infty)\right]\\ &= \mathbb{E}\left[\sum_j \big(-\Delta B_j^+(\infty)\big)U_j(\infty)\right]\\ &\le \sqrt{\mathbb{E}\left[\|\Delta\mathbf{B}^+(\infty)\|_2^2\right]}\cdot\sqrt{\mathbb{E}\left[\|\mathbf{U}(\infty)\|_2^2\right]}.\end{aligned}$$

Now each individual $U_j(\infty)$ cannot exceed the total capacity, so $\|\mathbf{U}(\infty)\|_2^2 = \sum_j U_j^2(\infty) \le C\sum_j U_j(\infty)$. Furthermore, by stationarity,

$$\mathbb{E}\left[\sum_j U_j(\infty)\right] = C - \mathbb{E}[\tilde{D}] = C - \Lambda = n\zeta.$$

Thus,

$$\mathbb{E}\left[\|\mathbf{U}(\infty)\|_2^2\right] \le \mathbb{E}\left[C\sum_j U_j(\infty)\right] = Cn\zeta.$$

By Corollary EC.2,

$$\sqrt{\mathbb{E}\left[\|\Delta\mathbf{B}^+(\infty)\|_2^2\right]} \le \frac{\sqrt{n}\sum_j\lambda_j^2 + 14n^{3/2}C^2 + 12n^{3/2}D_{\max}^2}{\eta} + \sqrt{n}(D_{\max}+C).$$

Thus,

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left[\left(\sum_j B_j^+(\infty)\right)\left(\sum_j U_j(\infty)\right)\right] &\le \sqrt{\mathbb{E}\left[\|\Delta\mathbf{B}^+(\infty)\|_2^2\right]}\cdot\sqrt{\mathbb{E}\left[\|\mathbf{U}(\infty)\|_2^2\right]}\\ &\le \left(\frac{\sqrt{n}\sum_j\lambda_j^2 + 14n^{3/2}C^2 + 12n^{3/2}D_{\max}^2}{\eta} + \sqrt{n}(D_{\max}+C)\right)\cdot\sqrt{Cn\zeta}\\ &= \sqrt{C\zeta}\cdot\left(\frac{n\sum_j\lambda_j^2 + 14n^2C^2 + 12n^2D_{\max}^2}{\eta} + n(D_{\max}+C)\right),\end{aligned}$$

and

$$\mathbb{E}\left[\left(\sum_j B_j^+(\infty)\right)\left(\sum_j U_j(\infty)\right)\right] \le \sqrt{C\zeta}\cdot\left(\frac{n^2\sum_j\lambda_j^2 + 14n^3C^2 + 12n^3D_{\max}^2}{\eta} + n^2(D_{\max}+C)\right). \tag{EC.18}$$

To complete the proof of Proposition 1, we plug the bounds in (EC.17) and (EC.18) into (EC.16), and get

$$\mathbb{E}\left[\sum_j B'_j(\infty)\right] \le \frac{\Sigma^2}{2n\zeta} + \frac{K'_1 + \eta K'_2}{\eta\sqrt{\zeta}}, \tag{EC.19}$$

where

$$K'_1 = \sqrt{C}(n\sum_j \lambda_j^2 + 14n^2C^2 + 12n^2D_{\max}^2), K'_2 = \sqrt{C}n(D_{\max} + C).$$

Note that by Assumption 1, K'_1 (and K'_2) can be upper-bounded by some constants $K_1(l,u)$ (and $K_2(l,u)$) respectively, and this concludes the proof. □

EC.2.2. Proofs of Proposition 2, Corollary 1 and Corollary 2

Proof of Proposition 2. Consider a discrete-time make-to-order system with a single plant of capacity C that produces only one type of product. In each time period t, demand $\sum_{j=1}^n D_j(t)$ arrives, where $D_j(t)$ is the amount of demand for product j in the original system. Let $\tilde{D}(t) = \sum_{j=1}^n D_j(t)$. Then, $\mathbb{E}[\tilde{D}(t)] = \Lambda < C$ and $Var[\tilde{D}(t)] = \sum_{j=1}^n \sigma_j^2 = \Sigma^2$. Let $\tilde{B}(t)$ be the backlog at the end of period t, then for all t,

$$\tilde{B}(t) = (\tilde{B}(t-1) + \tilde{D}(t) - C)^+.$$

Alternatively, we can write

$$\tilde{B}(t) = \tilde{B}(t-1) + \tilde{D}(t) - C + \tilde{U}(t), \tag{EC.20}$$

where $\tilde{U}(t)$ is the unused capacity in period t. For now let us note that $\tilde{U}(t)\tilde{B}(t) = 0$, since if there is positive unused capacity, i.e., $\tilde{U}(t) > 0$, then the backlog in the next time period must have been cleared, i.e., $\tilde{B}(t) = 0$. Consequently, $\tilde{U}(t)(\tilde{B}(t-1) + \tilde{D}(t) - C + \tilde{U}(t)) = 0$, or $\tilde{U}(t)(\tilde{B}(t-1) + \tilde{D}(t) - C) = -\tilde{U}^2(t)$ for all t.

Suppose that both the original and the single-plant system start empty. Let π be a greedy production policy that produces as much as possible each time. Then, using the recursions (10) and (EC.20), it can be shown that for each t, $\sum_{j=1}^n B_j^\pi(t) = \tilde{B}(t)$. Therefore,

$$\Gamma(\pi) = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \sum_{j=1}^n B_j^\pi(t) = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \tilde{B}(t).$$

We next bound $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \tilde{B}(t)$. First, note that $\Lambda < C$, so $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \tilde{B}(t)$ is finite. Furthermore, there exists a unique stationary distribution for the process $\tilde{B}(\cdot)$, and if we let $\tilde{B}(\infty)$ be a random variable with this stationary distribution, then all moments of $\tilde{B}(\infty)$ are finite, and

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \tilde{B}(t) = \mathbb{E}[\tilde{B}(\infty)].$$

We now show that $\frac{\Sigma^2}{2n\zeta} + \frac{C-\Lambda}{2} \geq \mathbb{E}[\tilde{B}(\infty)] \geq \frac{\Sigma^2+n^2\zeta^2}{2n\zeta} - \frac{1}{2}C$. To this end, consider the conditional drift term $\mathbb{E}[\tilde{B}^2(t+1) - \tilde{B}^2(t) \mid \tilde{B}(t)]$. We have

$$\begin{aligned}
\mathbb{E}[\tilde{B}^2(t+1) - \tilde{B}^2(t) \mid \tilde{B}(t)] &= \mathbb{E}[(\tilde{B}(t) + \tilde{D}(t+1) - C + \tilde{U}(t+1))^2 - \tilde{B}^2(t) \mid \tilde{B}(t)] \\
&= \mathbb{E}[(\tilde{D}(t+1) - C)^2 \mid \tilde{B}(t)] + 2\mathbb{E}[\tilde{D}(t+1) - C \mid \tilde{B}(t)]\tilde{B}(t) \\
&\quad -\mathbb{E}[\tilde{U}^2(t+1) \mid \tilde{B}(t)] \\
&= \Sigma^2 + (C-\Lambda)^2 - 2(C-\Lambda)\tilde{B}(t) - \mathbb{E}[\tilde{U}^2(t+1) \mid \tilde{B}(t)].
\end{aligned}$$

Now take expectation on both sides, over the stationary distribution of $\tilde{B}(\cdot)$. Then, by stationarity, the left-hand side is zero, and

$$0 = \Sigma^2 + n^2\zeta^2 - 2n\zeta\mathbb{E}[\tilde{B}(\infty)] - \mathbb{E}[\tilde{U}^2(\infty)] \leq \Sigma^2 + n^2\zeta^2 - 2n\zeta\mathbb{E}[\tilde{B}(\infty)].$$

Thus, we have $\frac{\Sigma^2}{2n\zeta} + \frac{C-\Lambda}{2} \geq \mathbb{E}[\tilde{B}(\infty)]$. Next, using $\tilde{B}(t) = \tilde{B}(t-1) + \tilde{D}(t) - C + \tilde{U}(t)$, we know that $\mathbb{E}[\tilde{U}(\infty)] = C - \Lambda = n\zeta$. Furthermore, unused capacity can never exceed total capacity, so $\tilde{U}(\infty) \leq C$. Thus, $\mathbb{E}[\tilde{U}^2(\infty)] \leq \mathbb{E}[C\tilde{U}(\infty)] = Cn\zeta$. This implies that

$$\begin{aligned}
\mathbb{E}[\tilde{B}(\infty)] &= \frac{1}{2n\zeta}\left(\Sigma^2 + n^2\zeta^2 - \mathbb{E}[\tilde{U}^2(\infty)]\right) \\
&\geq \frac{\Sigma^2}{2n\zeta} - \frac{C-n\zeta}{2}.
\end{aligned}$$

This completes the proof of Proposition 2. □

Proof of Corollary 1. Let $\mathcal{A}$ be a flexibility structure, and let π be a production policy that respects the structure $\mathcal{A}$. It is easy to see that $BL(\mathcal{A}) \geq BL(\mathcal{F}) \geq \frac{\Sigma^2}{2n\zeta} - \frac{C-n\zeta}{2}$. This proves the corollary. □

Proof of Corollary 2. Define $c = c_1$. By Remark 2, η of $\mathcal{LC}$ is equal to c. Combining this with Proposition 1 and Corollary 1, we have

$$\begin{aligned}
\frac{BL(\mathcal{LC})}{BL(\mathcal{F})} &\leq \left(\frac{\Sigma^2}{2n\zeta} - \frac{C-n\zeta}{2}\right)^{-1} \cdot \left(\frac{\Sigma^2}{2n\zeta} + \frac{K_1 + cK_2}{c\sqrt{\zeta}}\right) \\
&= 1 + \left(\frac{C-n\zeta}{2} + \frac{K_1 + cK_2}{c\sqrt{\zeta}}\right) \cdot \left(\frac{\Sigma^2}{2n\zeta} - \frac{C-n\zeta}{2}\right)^{-1}.
\end{aligned}$$

Now, pick a small $K^* > 0$ such that for all $\zeta \leq K^*$, $\frac{\Sigma^2}{2n\zeta} - \frac{C-n\zeta}{2} \geq \frac{\Sigma^2}{4n\zeta}$. We have then for all $\zeta \leq K^*$,

$$\begin{aligned}
\frac{BL(\mathcal{LC})}{BL(\mathcal{F})} &\leq 1 + \left(\frac{C-n\zeta}{2} + \frac{K_1 + cK_2}{c\sqrt{\zeta}}\right) \cdot \left(\frac{\Sigma^2}{2n\zeta} - \frac{C-n\zeta}{2}\right)^{-1} \\
&\leq 1 + \left(\frac{C-n\zeta}{2} + \frac{K_1 + cK_2}{c\sqrt{\zeta}}\right) \cdot \left(\frac{4n\zeta}{\Sigma^2}\right)
\end{aligned}$$

$$\leq 1+\Big(\frac{4n\sqrt{\zeta}}{\Sigma^2}\Big)\Big(\frac{\sqrt{\zeta}C}{2}+\frac{K_1+cK_2}{c}\Big),$$

where K_1 and K_2 are values specified in Proposition 1. Now, by Assumption 1, ζ, C, K_1 and K_2 are upper-bounded by a positive constant, while c, and Σ^2 are lower-bounded by a positive constant. Therefore, we have that there exists $K=K(l,u)>0$ such that

$$\frac{BL(\mathcal{LC})}{BL(\mathcal{F})}\leq 1+K\sqrt{\zeta}. \quad \square$$

EC.3. Relationship between Max-Weight and Max-Flow Policies

In this section, we discuss the relationship between the Max-Weight and Max-Flow policies. First, we formally introduce the definition of Max-Flow policies.

EC.3.1. Max-Flow Policies

In a nutshell, Max-Flow policies consist of all production policies that solve for a production schedule to *greedily* minimize the total backlog at the end of each time period. More precisely, under a Max-Flow policy, in each time period t, the production output $\mathbf{g}(t)$ at time t is an optimal solution of the optimization problem Opt-M defined below:

$$\min_{\mathbf{g}(t)\in R(\mathcal{A})}\sum_{j=1}^{n}b_j(t),\ \text{where } \mathbf{b}(t)=\big(\mathbf{b}(t-1)+\mathbf{d}(t)-\mathbf{g}(t)\big)^+. \tag{Opt-M}$$

Problem Opt-M is the same optimal policy of the one-period MTO model studied by Jordan and Graves (1995), with product demand $\mathbf{b}(t-1)+\mathbf{d}(t)$. As suggested in Jordan and Graves (1995), Opt-M can be solved as a max-flow problem (hence the name Max-Flow policy). To see that this is the case, recall the definition of constraints (6)–(8) of the production polytope $R(\mathcal{A})$, and we obtain the following equivalent optimization problem:

$$\begin{aligned}
&\min \sum_{j=1}^{n} b_j(t) && \text{(Flow-M)}\\
\text{s.t.}\quad &\sum_{i=1}^{m} f_{i,j}+b_j(t)=b_j(t-1)+d_j(t),\ \forall 1\leq j\leq n,\\
&\sum_{j=1}^{n} f_{i,j}\leq c_i,\ \forall 1\leq i\leq m,\\
&f_{i,j}=0,\forall(\mathcal{S}_i,\mathcal{T}_j)\notin\mathcal{A},\\
&\mathbf{b}(t)\in\mathbb{R}_+^n,\mathbf{f}\in\mathbb{R}_+^{mn}.
\end{aligned}$$

Note that if we write variables $b_j(t)$ in the objective of Flow-M as $b_j(t)=b_j(t-1)+d_j(t)-\sum_{i=1}^{m}f_{i,j}$, it is easy to see that Flow-M is equivalent to an optimization problem of maximizing the

objective $\sum_{j=1}^{n}\sum_{i=1}^{m} f_{i,j}$ under the appropriate constraints. This optimization problem is equivalent to a bipartite max-flow problem with graph $\mathcal{A}$, supply nodes $\{\mathcal{S}_1,...,\mathcal{S}_m\}$, and demand nodes $\{\mathcal{T}_1,...,\mathcal{T}_n\}$, where supply node $\mathcal{S}_i$ supplies up to c_i units of flow and demand node $\mathcal{T}_j$ receives up to $b_j(t-1)+d_j(t)$ units of flow. If $f^*_{i,j}$ is a max-flow solution, then for each $j\in\{1,2,\ldots,n\}$, by letting $b^*_j(t)=b_j(t-1)+d_j(t)-\sum_{i=1}^{m} f^*_{i,j}$, we have that $(\mathbf{b}^*(t),\mathbf{f}^*)$ is an optimal solution for Flow-M. An example of a bipartite max-flow problem equivalent to Flow-M is illustrated in Figure EC.1.

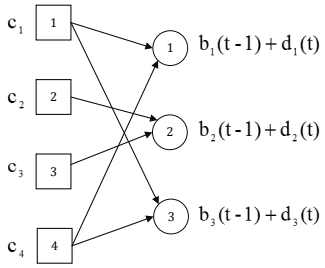

Figure EC.1 Max-flow Diagram for Flow-M with 4 plants and 3 products.

EC.3.2. Max-Weight Policies

Recall that a policy is a Max-Weight policy if it solves the optimization problem defined in Opt-MW at every time period t. Similar to the optimization problem Opt-M, we can expand $R(\mathcal{A})$ in Opt-MW to obtain a (weighted) max-flow formulation. In particular, Opt-MW is equivalent to:

$$\begin{aligned} & \max \sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}(b_j(t-1)+d_j(t)) && \text{(Flow-MW)}\\ \text{s.t.}\quad & \sum_{i=1}^{m} f_{i,j} \le b_j(t-1)+d_j(t),\\ & \sum_{j=1}^{n} f_{i,j} \le c_i,\ \forall 1\le i\le m\\ & f_{i,j}=0, \forall(\mathcal{S}_i,\mathcal{T}_j)\notin\mathcal{A},\ \mathbf{f}\in\mathbb{R}^{mn}_{+}. \end{aligned}$$

EC.3.3. Generalized Max-Flow Policies

Next, we present a proposition to show that a very general class of network flow optimization problems will lead to a Max-Flow policy. The proposition proves that the class of Max-Weight policies is contained in the class of Max-Flow policies as a special case.

PROPOSITION EC.3. *Consider the optimization problem*

$$\max \sum_{j=1}^{n} \Theta_j\left(\sum_{i=1}^{m} f_{i,j}\right) \qquad \text{(Flow-Monotone)}$$

$$
\begin{aligned}
s.t. \quad & \sum_{i=1}^{m} f_{i,j} \le b_j(t-1) + d_j(t), \\
& \sum_{j=1}^{n} f_{i,j} \le c_i, \forall 1 \le i \le m \\
& f_{i,j} = 0, \forall (\mathcal{S}_i, \mathcal{T}_j) \notin \mathcal{A} \\
& \mathbf{f} \in \mathbb{R}_+^{mn}.
\end{aligned}
$$

Suppose the Θ_j is a strictly increasing function for each $j = 1, \ldots, n$. Then, any optimal solution of Flow-Monotone is also an optimal solution of Flow-M.

Proof of Proposition EC.3. Let $\mathbf{f}^*$ be an optimal solution of Flow-Monotone. Suppose that there exists some augmenting path $P = (\mathcal{S}_{i_1}, \mathcal{T}_{i_2}, \mathcal{S}_{i_3}, \ldots \mathcal{S}_{i_{2k-1}}, \mathcal{T}_{i_{2k}})$ of $\mathbf{f}^*$, i.e., there exists $\epsilon > 0$ for which $\mathbf{f}^* + \epsilon(\sum_{l=1}^{k} \mathbf{e}^{i_{2l-1}, i_{2l}} - \sum_{l=1}^{k-1} \mathbf{e}^{i_{2l}, i_{2l+1}})$ is feasible, where for any pair $(i,j) \in \{(i_{2l-1}, i_{2l}) | l = 1, \ldots k\}$, $(i,j) \in \{(i_{2l}, i_{2l+1}) | l = 1, \ldots k-1\}$, $e_{i,j}^{i,j} = 1$ and $e_{i',j'}^{i,j} = 0, \forall (i',j') \neq (i,j)$.

Let $\mathbf{g} = \epsilon(\sum_{l=1}^{k} \mathbf{e}^{i_{2l-1}, i_{2l}} - \sum_{l=1}^{k-1} \mathbf{e}^{i_{2l}, i_{2l+1}})$. By definition, $\mathbf{f}^* + \mathbf{g}$ is feasible. Moreover, note that by the construction of $\mathbf{g}$,

$$
\begin{aligned}
& \sum_{i=1}^{m} f_{i,j} + g_{i,j} = \sum_{i=1}^{m} f_{i,j}, \forall j \neq i_2, \ldots, i_{2k} \\
& \sum_{i=1}^{m} f_{i,j} + g_{i,j} = \sum_{i=1}^{m} f_{i,j} + \epsilon - \epsilon = \sum_{i=1}^{m} f_{i,j}, \forall j \in \{i_2, \ldots, i_{2k-2}\} \\
& \sum_{i=1}^{m} f_{i,j} + g_{i,j} = \sum_{i=1}^{m} f_{i,j} + \epsilon, \text{ if } j = i_{2k}.
\end{aligned}
$$

Because Θ_j is a strictly increasing function for each $j = 1, \ldots, n$, we must have that

$$
\sum_{j=1}^{n} \Theta_j(\sum_{i=1}^{m} (f_{i,j} + g_{i,j})) > \sum_{j=1}^{n} \Theta_j(\sum_{i=1}^{m} (f_{i,j})),
$$

which contradicts the fact that $\mathbf{f}^*$ is optimal.

Therefore, there cannot exist any augmenting path of $\mathbf{f}^*$, that starts at plant node $\mathcal{S}_{i_1}$ and ends at product node $\mathcal{T}_{i_{2k}}$. Note that any augmenting path must start at a plant (supply) node and end at a product (demand) node and by the classical Ford-Fulkerson algorithm, $\mathbf{f}^*$ is an optimal solution of Flow-M. □

To see why Flow-MW returns a max-flow solution, consider the optimization problem Flow-MW', which has the same constraints as Flow-MW, and objective function $\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} w_j$ where

$$
\begin{aligned}
& w_j = b_j(t-1) + d_j(t), && \text{if } b_j(t-1) + d_j(t) > 0 \\
\text{and} \quad & w_j = 1, && \text{if } b_j(t-1) + d_j(t) = 0.
\end{aligned}
$$

Then, Flow-MW' is equivalent to Flow-MW, because if $b_j(t-1)+d_j(t)=0$, we have that $\sum_{j=1}^{n} f_{i,j} \leq b_j(t-1)+d_j(t)=0$.

Because Flow-MW' is a special instance of Flow-Monotone with $\Theta_j\left(\sum_{i=1}^{m} f_{i,j}\right)=w_j\sum_{i=1}^{m} f_{i,j}$, by Proposition EC.3, we immediately have that any optimal solution of Flow-MW' (and Flow-MW) is an optimal solution of Flow-M.

Next, we further analyze the optimization problems in the class of Flow-Monotone, with linear objectives. In particular, we present a result which states that if the objective function is in the form of $\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}w_j$, then the optimal solution is determined only by the ordering of the $w_1,\dots w_n$, and is independent of their absolute differences.

PROPOSITION EC.4. *Consider the optimization problem*

$$\begin{aligned} & \max \sum_{j=1}^{n}\sum_{i=1}^{m} w_j \cdot f_{i,j} && \text{(Flow}(\boldsymbol{w})\text{)} \\ s.t. \quad & \sum_{i=1}^{m} f_{i,j} \leq b_j(t-1)+d_j(t), \\ & \sum_{j=1}^{n} f_{i,j} \leq c_i, \forall 1 \leq i \leq m \\ & f_{i,j}=0, \forall (\mathcal{S}_i,\mathcal{T}_j) \notin \mathcal{A} \\ & \mathbf{f} \in \mathbb{R}_+^{mn}, \end{aligned}$$

where w_j is the linear weight for all of the flows that enter demand node $\mathcal{T}_j$. Let $\boldsymbol{w}^1$, $\boldsymbol{w}^2 \in \mathbb{R}_+^n$ be two strictly positive vectors where the entries have the same order, i.e.,

$$w_i^1 \leq w_j^1 \text{ if and only if } w_i^2 \leq w_j^2, \forall 1 \leq i,j \leq n.$$

Then, the set of optimal solutions for Flow($\boldsymbol{w}^1$) coincides the optimal solutions of Flow($\boldsymbol{w}^2$).

Proof of Proposition EC.4. Let $\mathbf{f}^1$ be an optimal solution of Flow($\boldsymbol{w}^1$). Suppose that $\mathbf{f}^1$ is not optimal for Flow($\boldsymbol{w}^2$), then there must exist some vector $\mathbf{g}$ such that $\mathbf{f}^1+\mathbf{g}$ is feasible for Flow($\boldsymbol{w}^2$), and

$$\sum_{j=1}^{n}\sum_{i=1}^{m} w_j^2 \cdot g_{i,j} > 0.$$

By the Flow Decomposition Theorem (see Theorem 3.5 in Ahuja et al. (1993) for details), we can always decompose $\mathbf{g}$ into a flow on path and cycles.

Suppose that $\mathbf{g}^C$ is a *cycle flow* on some cycle C with flow value ϵ. Note that for any product node $\mathcal{T}_j$ in C, we must have exactly two plant nodes, say $\mathcal{S}_{i_1}$ and $\mathcal{S}_{i_2}$, such that $(\mathcal{S}_{i_1},\mathcal{T}_j,\mathcal{S}_{i_2})$ is

directed path in C. This implies that $\epsilon = g_{i_1,j} = -g_{i_2,j}$, which in turn implies that $w_j^2 \cdot \sum_{i=1}^m \cdot g_{i,j} = 0$. Therefore, for any $\mathbf{g}^C$, we must have that

$$\sum_{j=1}^{n}\sum_{i=1}^{m} w_j^2 \cdot g_{i,j}^C = 0.$$

Therefore, we must have some vector $\mathbf{g}^P$ that is a *path flow* on some path P such that

$$\sum_{j=1}^{n}\sum_{i=1}^{m} w_j^2 \cdot g_{i,j}^P > 0. \tag{EC.21}$$

Suppose that $\mathbf{g}^P$ is one such path flow with flow value ϵ. Note that for any product node $\mathcal{T}_j$ in P, there exists integers i_1, i_2 such that we either have $(\mathcal{T}_j, \mathcal{S}_{i_2})$ to be the first arc in path P, or $(\mathcal{S}_{i_1}, \mathcal{T}_j)$ to be the last arc in path P, or $(\mathcal{S}_{i_1}, \mathcal{T}_j, \mathcal{S}_{i_2})$ to be a directed path in P.

If $(\mathcal{T}_j, \mathcal{S}_{i_2})$ is the first arc in P, then

$$\sum_{i=1}^{m} w_j^2 \cdot g_{i,j}^P = w_j^2 g_{i_2,j}^P = -w_j^2 \epsilon.$$

If $(\mathcal{S}_{i_1}, \mathcal{T}_j)$ is the last arc in P, then

$$\sum_{i=1}^{m} w_j^2 \cdot g_{i,j}^P = w_j^2 g_{i_2,j}^P = w_j^2 \epsilon.$$

And finally, if $(\mathcal{S}_{i_1}, \mathcal{T}_j, \mathcal{S}_{i_2})$ is a directed path in P, then

$$\sum_{i=1}^{m} w_j^2 \cdot g_{i,j}^P = 0.$$

By (EC.21), and the equations above, we must have some product node $\mathcal{T}_{j_2}$ such that it is the last node in P. If the first node in P is a plant node, note that $\mathbf{f}^1 + \mathbf{g}^P$ is feasible and by equations above, we have that

$$\sum_{j=1}^{n}\sum_{i=1}^{m} w_j^1 \cdot g_{i,j}^P = w_{j_2}^1 \epsilon > 0,$$

which is a contradiction to the optimality of $\mathbf{f}^1$ of Flow($\boldsymbol{w}^1$).

Thus, the first node in P must be a product node. Let the product node be $\mathcal{T}_{j_1}$, then we have that

$$0 < \sum_{j=1}^{n}\sum_{i=1}^{m} w_j^2 \cdot g_{i,j}^P = (w_{j_2}^2 - w_{j_1}^2)\epsilon \implies w_{j_2}^2 > w_{j_1}^2.$$

But this implies that $w^1_{j_2} > w^1_{j_1}$. And because

$$\sum_{j=1}^{n}\sum_{i=1}^{m} w^1_j \cdot g^P_{i,j} = (w^1_{j_2} - w^1_{j_1})\epsilon,$$

we have that $\mathbf{f}^1 + \mathbf{g}^P$ is a strictly better feasible solution for Flow($\boldsymbol{w}^1$), which results in a contradiction. □

Since Flow($\boldsymbol{w}$) belongs to the class of Flow-Monotone, it is a max-flow solution by Proposition EC.3. Therefore, Proposition EC.4 suggests that any optimization of the form Flow($\boldsymbol{w}$) is essentially an optimal solution of Flow-M that prioritizes products with higher linear weights.

Another interesting implication of Proposition EC.4 concerns the generality of the Max-Weight policy. Suppose that $f(\cdot)$ is a strictly increasing function, and the factors $b_j(t-1) + d_j(t)$ in the objective of Flow-MW are replaced by $f(b_j(t-1) + d_j(t))$. By Proposition EC.4, this new optimization problem has the same set of optimal solutions as Flow-MW. In particular, this implies that in our model, the well-studied Max-Weight-α policies (where $f(x) = x^\alpha$, $\alpha > 0$, see Keslassy and McKeown (2001), Shah and Wischik (2012) for more details) all coincide with the Max-Weight policy, a fact that is not necessarily true in other queueing models.

EC.4. Proofs in §5

Proof of Lemma 4. We prove the lemma using backward induction. $t = T+1$ is trivially true. Suppose the statement is true for $t = K$. For any b_1, b_2 where $b_1 \geq 1$, let $p^K_1(\mathbf{b}, \mathbf{d})$, $p^K_2(\mathbf{b}, \mathbf{d})$ be the optimal production of product 1 and 2 at time t given that $\mathbf{b}$ and $\mathbf{d}$ are the backlog and demand at time $t = K$, respectively. Note that by induction hypothesis, we must have

$$p^K_1(\mathbf{b}, \mathbf{d}) = \min\{c_1, b_1 + d_1\} \tag{EC.22}$$

$$p^K_2(\mathbf{b}, \mathbf{d}) = \min\{c_2 + (c_1 - b_1 - d_1)^+, b_2 + d_2\}. \tag{EC.23}$$

By Equations (EC.22) and (EC.23), we can define $b^{K+1}_1(\mathbf{b}, \mathbf{d})$, $b^{K+1}_2(\mathbf{b}, \mathbf{d})$, the backlogs in time period $K+1$, as follows.

$$b^{K+1}_1(\mathbf{b}, \mathbf{d}) = \max\{b_1 + d_1 - c_1, 0\} \tag{EC.24}$$

$$b^{K+1}_2(\mathbf{b}, \mathbf{d}) = \max\{b_2 + d_2 - c_2 - (c_1 - b_1 - d_1)^+, 0\}. \tag{EC.25}$$

Let $\mathbf{b}' = [b_1 - 1, b_2 + 1]$. First, it is simple to check that for any $\mathbf{d}$, if $b^{K+1}_1(\mathbf{b}, \mathbf{d}) \geq 1$, then we must have that

$$b^{K+1}_1(\mathbf{b}, \mathbf{d}) = b^{K+1}_1(\mathbf{b}', \mathbf{d}) - 1,$$

$$\text{and} \qquad b_2^{K+1}(\mathbf{b},\mathbf{d}) \geq b_2^{K+1}(\mathbf{b}',\mathbf{d}) + 1.$$

Therefore, if $b_1^{K+1}(\mathbf{b},\mathbf{d}) \geq 1$, we must have that for any $\mathbf{d}$,

$$b_1^{K+1}(\mathbf{b},\mathbf{d}) + b_2^{K+1}(\mathbf{b},\mathbf{d}) \geq b_1^{K+1}(\mathbf{b}',\mathbf{d}) + b_2^{K+1}(\mathbf{b}',\mathbf{d}),$$
$$\text{and} \qquad J^{K+1}(b_1^{K+1}(\mathbf{b},\mathbf{d}), b_2^{K+1}(\mathbf{b},\mathbf{d})) \geq J^{K+1}(b_1^{K+1}(\mathbf{b}',\mathbf{d}), b_2^{K+1}(\mathbf{b}',\mathbf{d})).$$

This implies that

$$J^K(b_1,b_2) \geq J^K(b_1-1, b_2+1).$$

Next, if $b_1^{K+1}(\mathbf{b},\mathbf{d}) = 0$, and $b_2^{K+1}(\mathbf{b},\mathbf{d}) \geq 1$, then we must have that

$$b_1^{K+1}(\mathbf{b},\mathbf{d}) = b_1^{K+1}(\mathbf{b}',\mathbf{d}) = 0,$$
$$\text{and} \qquad b_2^{K+1}(\mathbf{b},\mathbf{d}) = b_2^{K+1}(\mathbf{b}',\mathbf{d}),$$

which implies that

$$J^K(b_1,b_2) = J^K(b_1-1, b_2+1).$$

Therefore, in either case, we must have that

$$J^K(b_1,b_2) \geq J^K(b_1-1, b_2+1),$$

and the proof is done by induction. □

Proof of Corollary 4. Applying Proposition 6, we have

$$\begin{aligned}
\frac{BL(\mathcal{A})}{BL(\mathcal{F})} &\geq \left(\frac{\Sigma_\Omega^2}{2(\eta+|\Omega|\zeta)} + \frac{\Sigma_{\Omega^c}^2}{2n\zeta} - \frac{\Lambda + \sum_{\mathcal{T}_j \in \Omega} \lambda_j}{2}\right) \cdot \left(\frac{\Sigma^2}{2n\zeta} + \frac{C-\Lambda}{2}\right)^{-1} \\
&= 1 + \left(\frac{\Sigma_\Omega^2}{2(\eta+|\Omega|\zeta)} - \frac{\Sigma_\Omega^2}{2n\zeta} - \frac{\Lambda + \sum_{\mathcal{T}_j \in \Omega} \lambda_j}{2} - \frac{C-\Lambda}{2}\right) \cdot \left(\frac{\Sigma^2}{2n\zeta} + \frac{C-\Lambda}{2}\right)^{-1} \\
&\geq 1 + \left(\frac{\Sigma_\Omega^2}{2(\eta+|\Omega|\zeta)} - \frac{\Sigma_\Omega^2}{2n\zeta} - C\right) \cdot \left(\frac{\Sigma^2}{2n\zeta} + \frac{n\zeta}{2}\right)^{-1} \\
&\geq 1 + \left(\frac{\Sigma_\Omega^2}{2(n-\alpha)\zeta} - \frac{\Sigma_\Omega^2}{2n\zeta} - C\right) \cdot \left(\frac{\Sigma^2}{2n\zeta} + \frac{n\zeta}{2}\right)^{-1} \\
&\geq 1 + \left(\frac{\Sigma_\Omega^2}{2(n-\alpha)\zeta} - \frac{\Sigma_\Omega^2}{2n\zeta} - C\right) \cdot \frac{n\zeta}{\Sigma^2} \\
&= 1 + \left(\frac{\alpha\Sigma_\Omega^2}{2n(n-\alpha)\zeta} - C\right) \cdot \frac{n\zeta}{\Sigma^2} \\
&\geq 1 + \frac{\alpha\Sigma_\Omega^2}{4(n-\alpha)\Sigma^2} \\
&\geq 1 + \frac{\alpha l}{4n(n-\alpha)u^2}.
\end{aligned}$$

The third inequality above follows from the facts that $\eta < (1-\alpha)\zeta$, $|\Omega| \le n-1$; the fourth inequality follows from $\zeta \le \frac{l}{n} \le \frac{\sqrt{\Sigma^2}}{n}$ thus implying that $\frac{n\zeta}{2} \le \frac{\Sigma^2}{2n\zeta}$; and the fifth inequality above follows from the fact that $\zeta \le \frac{\alpha l^2}{4mn(n-\alpha)u} \le \frac{\alpha \Sigma_\Omega^2}{4Cn(n-\alpha)}$. □

EC.5. Numerical Results for §6

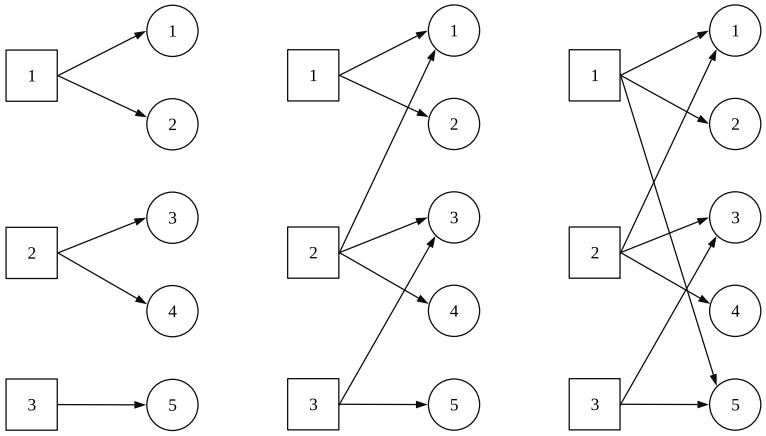

Figure EC.2 Structures for §6.1.2: the 3 by 5 systems from left to right: Dedicated, $\mathcal{C}^-$, $\mathcal{C}$.

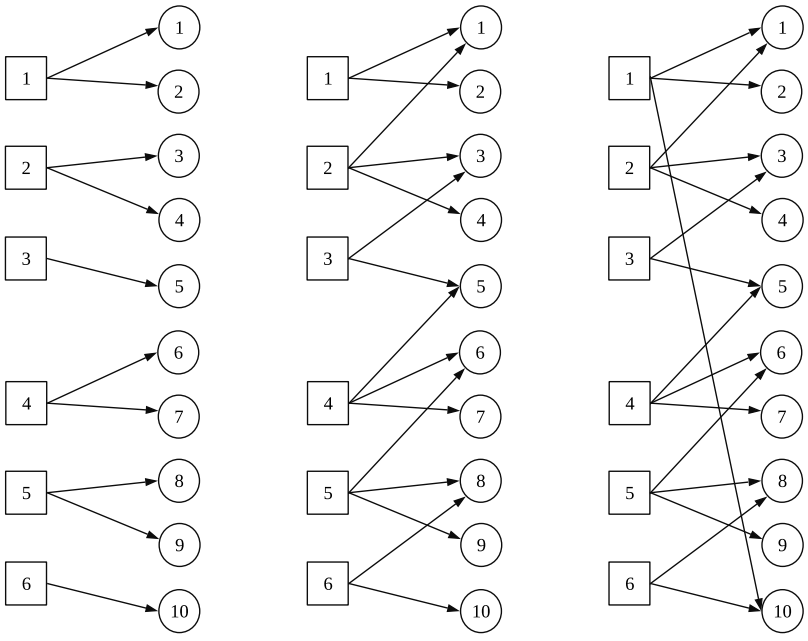

Figure EC.3 Structures for §6.1.2: the 6 by 10 systems from left to right: Dedicated, $\mathcal{C}^-$, $\mathcal{C}$.

ρ	cv	$R(\mathcal{LC})$	$\Delta(\mathcal{LC})$	$B(\mathcal{LC})$	SE%	$R(\mathcal{LC})$	$\Delta(\mathcal{LC})$	$B(\mathcal{LC})$	SE%
		$n=5$				$n=10$			
0.8	0.3	100.0%	100.0%	0.7	1.2%	112.5%	100.0%	0.1	3.0%
0.8	0.4	100.9%	99.9%	3.1	0.8%	145.2%	99.6%	1.3	1.3%
0.8	0.5	103.2%	99.8%	7.2	0.7%	182.0%	98.8%	5.1	0.9%
0.9	0.3	100.3%	100.0%	12.0	0.6%	109.1%	99.7%	7.6	0.7%
0.9	0.4	101.7%	99.8%	28.6	0.6%	131.6%	98.7%	25.1	0.6%
0.9	0.5	104.5%	99.4%	50.5	0.7%	159.7%	97.1%	57.3	0.6%
0.95	0.3	100.4%	100.0%	50.0	0.7%	108.1%	99.6%	42.9	0.6%
0.95	0.4	101.6%	99.8%	99.6	0.8%	124.9%	98.5%	104.5	0.7%
0.95	0.5	103.4%	99.5%	157.5	0.9%	142.8%	97.0%	196.6	0.8%
0.975	0.3	100.3%	100.0%	137.2	0.9%	106.5%	99.6%	131.4	0.8%
0.975	0.4	101.1%	99.8%	254.4	1.0%	117.8%	98.5%	272.9	0.9%
0.975	0.5	102.1%	99.6%	380.7	1.0%	128.1%	97.0%	462.7	0.9%
0.9875	0.3	100.2%	100.0%	298.1	1.0%	104.5%	99.5%	311.6	1.0%
0.9875	0.4	100.7%	99.8%	496.1	0.9%	111.2%	98.3%	590.3	1.0%
0.9875	0.5	101.3%	99.5%	705.6	0.9%	117.0%	96.9%	907.6	1.0%

Table EC.1 Performance of the long chain with $n=5$ and 10

ρ	cv	$R(\mathcal{LC})$	$\Delta(\mathcal{LC})$	$B(\mathcal{LC})$	SE%	$R(\mathcal{LC})$	$\Delta(\mathcal{LC}))$	$B(\mathcal{LC})$	SE%
		$n=15$				$n=20$			
0.8	0.3	335.6%	100.0%	0.0	8.0%	1570.3%	99.9%	0.1	18.7%
0.8	0.4	453.7%	99.4%	1.3	2.3%	1678.0%	99.3%	1.6	4.0%
0.8	0.5	522.5%	98.2%	6.5	1.4%	1500.4%	98.0%	8.5	2.0%
0.9	0.3	152.9%	99.4%	6.5	0.8%	260.9%	99.1%	7.3	0.9%
0.9	0.4	222.5%	97.6%	30.7	0.7%	378.4%	96.9%	39.6	0.7%
0.9	0.5	282.5%	95.4%	78.8	0.6%	472.1%	94.3%	104.3	0.6%
0.95	0.3	134.6%	99.0%	45.2	0.7%	184.1%	98.4%	52.6	0.6%
0.95	0.4	181.8%	97.0%	134.7	0.7%	262.8%	95.9%	173.2	0.6%
0.95	0.5	222.0%	94.8%	269.7	0.7%	318.1%	93.5%	356.9	0.7%
0.975	0.3	127.1%	98.8%	144.0	0.8%	161.6%	98.1%	169.6	0.7%
0.975	0.4	158.1%	96.7%	345.0	0.9%	212.5%	95.4%	443.4	0.8%
0.975	0.5	179.6%	94.6%	610.5	0.9%	244.6%	93.0%	788.1	0.8%
0.9875	0.3	119.2%	98.6%	341.3	1.0%	143.7%	97.6%	401.0	0.9%
0.9875	0.4	137.6%	96.3%	722.5	1.0%	173.2%	94.6%	900.4	1.0%
0.9875	0.5	148.2%	94.1%	1171.2	1.0%	190.3%	92.1%	1450.7	1.0%

Table EC.2 Performance of the long chain with $n=15$ and 20

ρ	cv	$R(\mathcal{LC}^-)$	$\Delta(\mathcal{LC}^-)$	$R(\mathcal{LC}^-)$	$\Delta(\mathcal{LC}^-)$	$R(\mathcal{LC}^-)$	$\Delta(\mathcal{LC}^-)$	$R(\mathcal{LC}^-)$	$\Delta(\mathcal{LC}^-)$
		$n=5$		$n=10$		$n=15$		$n=20$	
0.8	0.3	809%	80%	5704%	88%	40662%	92%	169102%	94%
0.8	0.4	647%	70%	2412%	82%	7681%	87%	23141%	90%
0.8	0.5	586%	64%	1739%	76%	4094%	83%	9367%	86%
0.9	0.3	417%	70%	1295%	66%	2271%	75%	3556%	80%
0.9	0.4	361%	69%	964%	65%	1852%	66%	2571%	72%
0.9	0.5	333%	69%	814%	65%	1512%	64%	2154%	69%
0.95	0.3	298%	81%	683%	73%	1175%	70%	1810%	68%
0.95	0.4	273%	78%	587%	71%	961%	68%	1422%	67%
0.95	0.5	261%	76%	532%	70%	866%	67%	1232%	66%
0.975	0.3	265%	80%	523%	72%	817%	69%	1136%	68%
0.975	0.4	235%	77%	458%	69%	703%	66%	961%	65%
0.975	0.5	218%	74%	412%	67%	630%	64%	871%	63%
0.9875	0.3	238%	74%	420%	65%	615%	62%	819%	61%
0.9875	0.4	202%	73%	345%	63%	503%	60%	664%	58%
0.9875	0.5	185%	72%	308%	62%	442%	58%	591%	57%

Table EC.3 Performance of long chain less arc $(n,1)$

ρ	cv	$R(\mathcal{C})$	$\Delta(\mathcal{C})$	$B(\mathcal{C})$	$R(\mathcal{C}^-)$	$\Delta(\mathcal{C}^-)$	SE%
0.8	0.3	100.3%	100.0%	0.53	569.8%	60.2%	1.1%
0.8	0.4	101.6%	99.8%	2.24	393.7%	57.7%	0.7%
0.8	0.5	104.2%	99.2%	5.17	337.7%	55.8%	0.7%
0.9	0.3	101.1%	99.8%	8.25	231.8%	71.5%	0.6%
0.9	0.4	104.0%	99.0%	19.92	223.6%	67.4%	0.6%
0.9	0.5	107.9%	97.7%	34.99	222.1%	64.6%	0.7%
0.95	0.3	102.1%	99.7%	34.13	179.2%	89.7%	0.8%
0.95	0.4	105.2%	99.1%	68.75	181.9%	85.1%	0.8%
0.95	0.5	108.4%	98.1%	111.22	177.7%	82.5%	0.8%
0.975	0.3	102.4%	99.7%	92.04	153.5%	93.7%	1.0%
0.975	0.4	104.6%	99.1%	171.09	161.0%	88.0%	1.0%
0.975	0.5	106.4%	98.2%	262.83	158.2%	84.1%	1.0%
0.9875	0.3	102.1%	99.6%	199.74	137.1%	93.1%	1.0%
0.9875	0.4	103.4%	99.0%	337.47	141.4%	87.6%	0.9%
0.9875	0.5	104.6%	98.2%	464.54	143.4%	82.9%	0.9%

Table EC.4 Performance of $\mathcal{C}^-$ and $\mathcal{C}$ in the 3 by 5 system

ρ	cv	$R(\mathcal{C})$	$\Delta(\mathcal{C})$	$B(\mathcal{C})$	$R(\mathcal{C}^-)$	$\Delta(\mathcal{C}^-)$	SE%
0.8	0.3	118.6%	99.9%	0.11	3520.2%	77.1%	2.6%
0.8	0.4	151.3%	99.0%	1.04	1421.9%	73.4%	1.2%
0.8	0.5	176.3%	97.4%	3.67	973.3%	70.2%	0.9%
0.9	0.3	113.5%	99.3%	5.50	559.5%	74.8%	0.7%
0.9	0.4	134.9%	97.3%	17.85	527.4%	67.1%	0.6%
0.9	0.5	158.6%	94.7%	38.08	505.4%	63.5%	0.6%
0.95	0.3	113.3%	99.4%	30.44	425.1%	84.4%	0.7%
0.95	0.4	131.2%	97.8%	72.22	382.6%	80.5%	0.7%
0.95	0.5	147.5%	95.8%	132.60	354.1%	77.3%	0.8%
0.975	0.3	112.9%	99.3%	91.74	381.6%	85.8%	0.8%
0.975	0.4	125.3%	97.9%	191.95	330.4%	80.9%	0.9%
0.975	0.5	135.8%	96.0%	317.60	304.3%	77.0%	0.9%
0.9875	0.3	110.8%	99.1%	212.96	377.7%	76.9%	1.0%
0.9875	0.4	118.2%	97.5%	405.35	286.7%	74.3%	1.0%
0.9875	0.5	123.7%	95.5%	615.38	250.2%	71.8%	1.0%

Table EC.5 Performance of $\mathcal{C}^-$ and $\mathcal{C}$ in the 6 by 10 system

ρ	cv	$R(\mathcal{C})$	$\Delta(\mathcal{C})$	$B(\mathcal{C})$	$R(\mathcal{C}^-)$	$\Delta(\mathcal{C}^-)$	SE%
0.8	0.3	386.3%	99.8%	0.06	19390.1%	84.6%	5.9%
0.8	0.4	442.7%	98.4%	1.07	4215.1%	80.9%	2.1%
0.8	0.5	471.9%	96.2%	4.57	2278.7%	77.5%	1.2%
0.9	0.3	165.0%	98.5%	5.16	966.2%	80.2%	0.8%
0.9	0.4	220.6%	95.6%	21.36	834.0%	73.3%	0.7%
0.9	0.5	267.9%	92.3%	51.27	814.2%	67.2%	0.6%
0.95	0.3	144.1%	98.8%	32.70	727.3%	83.2%	0.7%
0.95	0.4	184.0%	96.6%	91.68	609.7%	79.3%	0.7%
0.95	0.5	218.1%	93.9%	178.46	560.0%	76.3%	0.7%
0.975	0.3	135.4%	98.9%	101.74	601.0%	84.7%	0.8%
0.975	0.4	162.9%	96.8%	236.50	504.8%	79.3%	0.9%
0.975	0.5	182.4%	94.2%	415.39	455.6%	74.8%	0.9%
0.9875	0.3	127.5%	98.6%	239.23	570.6%	75.2%	1.0%
0.9875	0.4	144.2%	96.1%	493.41	421.0%	71.7%	1.0%
0.9875	0.5	154.6%	93.5%	774.19	359.8%	69.2%	1.0%

Table EC.6 Performance of $\mathcal{C}^-$ and $\mathcal{C}$ in the 9 by 15 system

n	ρ	$R(\mathcal{LC})$	$\Delta(\mathcal{LC})$	$R(\mathcal{LC}^-)$	$\Delta(\mathcal{LC}^-)$	SE%
5	0.8	179.2%	87.6%	356.8%	59.8%	1.1%
5	0.9	137.0%	93.4%	274.9%	68.9%	1.5%
5	0.95	119.1%	97.3%	228.3%	81.8%	2.1%
5	0.975	108.8%	98.6%	187.2%	85.9%	2.7%
5	0.9875	104.5%	98.9%	159.7%	84.9%	3.0%
10	0.8	456.9%	80.9%	738.1%	65.8%	0.9%
10	0.9	271.3%	87.0%	493.7%	70.2%	1.3%
10	0.95	189.5%	93.2%	390.3%	78.1%	1.9%
10	0.975	146.1%	95.8%	310.2%	80.8%	2.6%
10	0.9875	124.0%	96.7%	250.4%	79.2%	2.9%
15	0.8	890.1%	78.7%	1313.7%	67.3%	0.9%
15	0.9	447.3%	84.2%	739.8%	70.9%	1.1%
15	0.95	290.5%	90.8%	563.8%	77.5%	1.6%
15	0.975	202.0%	93.9%	444.8%	79.2%	2.6%
15	0.9875	154.9%	95.0%	356.9%	76.7%	2.7%

Table EC.7 Performance of $\mathcal{LC}$ and $\mathcal{LC}^-$ in a continuous parallel system

n	wip	$R(\mathcal{LC})$	$\Delta(\mathcal{LC})$	$R(\mathcal{LC}^-)$	$\Delta(\mathcal{LC}^-)$	SE%
5	5.00	82.6%	61.0%	64.8%	20.9%	0.03%
5	10.00	95.5%	84.3%	77.4%	21.5%	0.02%
5	25.00	100.0%	99.7%	88.4%	18.2%	0.02%
10	10.00	75.4%	48.2%	65.2%	26.7%	0.03%
10	20.00	91.1%	71.3%	79.4%	33.9%	0.03%
10	50.00	99.3%	95.2%	88.5%	25.7%	0.03%
15	15.00	72.8%	43.8%	65.7%	29.0%	0.03%
15	30.00	89.4%	66.8%	80.6%	39.3%	0.03%
15	75.00	98.6%	91.5%	89.7%	35.5%	0.03%

Table EC.8 Performance of $\mathcal{LC}$ and $\mathcal{LC}^-$ in a serial production line