Characterizing the Performance of Process Flexibility Structures

O. Zeynep Akşin [*]   Fikri Karaesmen [†]

July 2004, Revisions, September 2005; January 2006

Abstract

The objective is to identify preferred flexibility structures in service or manufacturing systems, when demand is random and capacity is finite. Considering a network flow type model as the basis of the analysis, general structural properties of flexibility design pertaining to the marginal values of flexibility and capacity are identified.

Keywords: flexibility; network flow problem; service systems; call centers; cross-training

1 Introduction

This paper considers service systems with multi-departmental structures having possibly multi-skill servers that treat several types of service requests. In any such system, it is possible to have a different mix of skill sets with a different number of servers belonging to each skill set. It is well known that more flexibility leads to better operational performance. However given that there are costs associated with creating and maintaining this flexibility, and difficulties managing the resulting more complex system, it is desirable to understand the value of this flexibility in more depth. This paper will focus on providing a better understanding of the relationship between different flexibility structures and value. The following questions are relevant in this setting: How many skills should servers have (*how much flexibility*)? What are the ideal skill-sets for those that are cross-trained (*what type of flexibility*)? How should these skill-sets be formed

[*] Graduate School of Business, Koç University, zaksin@ku.edu.tr

[†] Corresponding author: Deparment of Industrial Engineering, Koç University, Istanbul, 34450, Turkey, fkaraesmen@ku.edu.tr

in a multi-departmental structure, where each server has a primary skill and some secondary skills (*where*)? This set of questions motivate our research and will be labeled as the flexibility design problem in the ensuing analysis. We provide guidelines that will be useful in addressing such a flexibility design problem.

A well known application of this flexibility design problem in a manufacturing setting is the problem studied in [9]. In this setting, the departments are different plants or production lines, while the customer types represent different products to be produced in these production facilities. Process flexibility constitutes the ability of producing a product in multiple plants or production lines. The model that we study in Section 3 is identical to the one in [9]. Using this model as a basis, we formalize some results pertaining to the performance of different flexible structures that were observed numerically in [9].

The remaining parts of this paper are organized as follows. Related literature is reviewed in the next section. Section 3 introduces the model and the problem. The results on flexibility/capacity interactions are presented in Section 4. Section 5 presents the results on the diminishing returns property of flexibility. Finally in Section 6 results pertaining to balance in flexibility structures are stated.

2 Literature Review

The importance of flexibility in service delivery is well known. A significant source of service delivery process flexibility comes from the use of cross-trained servers. While the practice itself is widespread, there is little formal evaluation of the value of this type of practice from an operations standpoint. In [13] trade-offs between capacity and quality for cross-trained workers in service systems are considered. Numerous studies in manufacturing have looked at the case of flexible workers and their impact on performance in terms of operational measures like throughput. Most of these studies analyze specific work-sharing schemes in queueing network models ([16], and the references therein). Karaesmen et al. [10] investigate flexibility in the context of field service design. These papers assess the value of certain workforce flexibility practices in given settings, however do not tackle the broader question of designing the type of flexibility in these systems.

More generally, the benefits and design of flexibility in operations have been studied extensively ([2], [14]). An important stream consists of papers that address the capacity investment problem in the presence of flexible resources ([3], [17], [12]). These papers assume a certain form of flexibility and then explore the question of the ideal level of this flexibility and how it relates to value under uncertain demand.

In this paper, we consider capacity to be fixed and explore the relationship between different flexibility structures and value, without explicitly addressing the optimal capacity issue. In this regard, our analysis parallels that in [9]. Focusing on process flexibility, Jordan and Graves explore the problem of assigning multiple products to multiple plants, where the flexibility of the plants determines which products they can handle. The authors illustrate that well designed limited flexibility is almost as good as full flexibility. To address the question of where flexibility should be added in a system, the authors define a *chain* structure as a group of directly or indirectly connected group of products and plants. It is shown that a structure that enables the formation of fewer long chains is superior to one with multiple short chains. The principles are illustrated with simulations. A similar analysis is performed for multi-stage systems in [5] and a flexibility measure is developed for such systems. Hopp et al. [7] explore the benefits of chaining in the context of cross-training for production lines. Gurumurthi and Benjaafar [6] present a numerical investigation of the benefits of chaining based on a queueing model. Finally, [8] develops a *structural flexibility index* that quantifies the structural flexibility in production and service systems, and allows a ranking based on performance of these systems.

The objective of this paper is to represent flexibility structures through a network flow model as in [9] and formalize earlier observations on flexibility using this model. By doing this, we will establish certain flexibility design principles for service and manufacturing systems. To our knowledge, this is the only paper providing analytical comparison results on flexibility structure in the literature. In particular, we focus on three sets of design issues: flexibility/capacity interactions, diminishing returns to increased flexibility, and the value of balance in flexibility structures.

3 Modeling Process Flexibility

Consider a service system with multiple customer types. Customer types differ in terms of their service requirements. Servers specialize by customer type, but can be flexible with overlapping skill sets, allowing them to treat customer requests from different types. The service system can be represented as a directed graph $G = (N, A)$ with a set of nodes N, of which one is a source

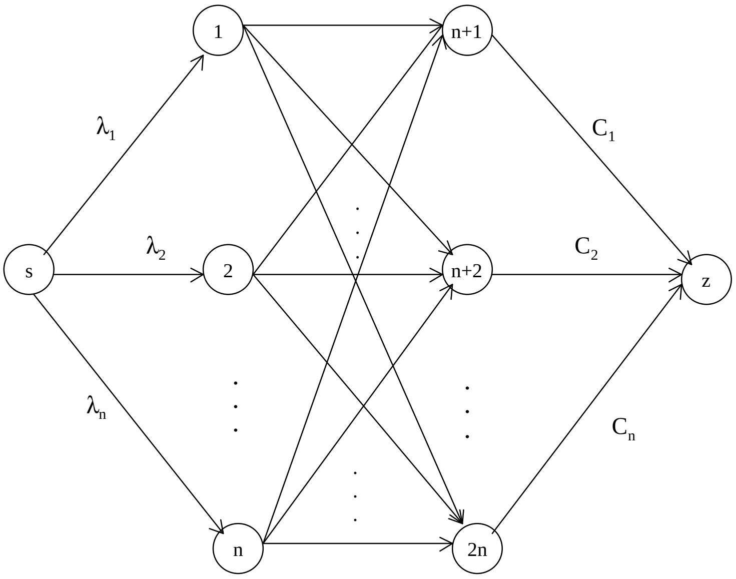

Figure 1: An n class service system with full flexibility

node and one a sink node, and a set of arcs A whose elements are ordered pairs of distinct nodes. Some standard definitions are useful to formalize the description of this network. A directed arc (i, j) emanating from node i is said to have tail i, terminating in node j known as the head of the arc. For an arc $(i, j) \in A$, the node j is said to be adjacent to node i. The node adjacency list $A(i)$ is the set of adjacent nodes, $A(i) = \{j \in N : (i, j) \in A\}$. The indegree of a node is the number of incoming arcs of that node and its outdegree is the number of outgoing arcs.

An instance of a network that represents the service system is depicted in Figure 3. This graph illustrates a system with n customer types given by the set of nodes $I = \{1, 2, ..., n\}$, served by servers in n departments, given by the set of nodes $J = \{n+1, n+2, ..., 2n\}$. Note that since servers are assumed to be organized by their primary skills, the number of customer types is equal to the number of departments. The case where the number of customer types is larger than the number of departments can also be treated within this framework, where the additional classes can be served by dummy departments with no servers in them. The arcs emanating from the source node s and terminating in nodes $i \in I$ represent the service demand, and have capacity given by the demand vector $\lambda = (\lambda_1, ..., \lambda_n)$. This vector represents the realization of demand for a given period. The arcs emanating from nodes $j \in J$ and terminating in the sink node z represent the capacity of each department. These arcs have a capacity given by the vector $\mathbf{C} = (C_1, ..., C_n)$. The arcs (i, j) with $i \in I$ and $j \in J$ represent the flexibility of the system. Whenever a customer of type $i \in I$ can be served by a server of type $j \in J$, an arc (i, j) with infinite capacity is added to the network. The network in Figure 3 illustrates a case where all customers can be treated by all servers, i.e. where the system has full flexibility. In a system with n departments, full flexibility implies that each node $i \in I$ has outdegree equal to n. In general, the outdegree of node $i \in I$ represents the number of possible routings for customers of type i, and the indegree of a node $j \in J$ represents the number of skills a server of type j has. Assuming that each customer request of type $i \in I$ is worth r to the system, the problem of maximizing the value generated by a given configuration for a demand realization $\boldsymbol{\lambda}$ is equivalent to the maximum flow problem for this network. We refer to the maximal flow as the throughput and denote it by $T(\boldsymbol{\lambda}, \mathbf{C})$. A closely related problem, first studied in [9], considers a random demand vector $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)$ where the performance measure of interest is the expected throughput $E[T(\boldsymbol{\lambda}, \mathbf{C})]$. In this case, a maximum flow problem is solved for each realization of the random demand vector $\boldsymbol{\lambda}$ and the expectation is taken over all realizations of $\boldsymbol{\lambda}$. Jordan and Graves also discuss the relevance of the expected throughput maximization objective and relate it to a number of other possible objectives.

As in [9], the emphasis will be on a special class of processing network which is defined below:

Definition 1 *An arc (i, j) is flexible if $i \in I$ and $j \in J$. The flexibility of a network is defined as the set of arcs $\mathcal{F}$ between the node set I and the node set J. All arcs in $\mathcal{F}$ are in parallel (i.e. there is no directed cycle in the graph in which both arcs have the same direction).*

Definition 2 *A symmetric network is defined as a network where i) every customer type can be processed by the same number of server types (departments), i.e. each node $i \in I$ has the same outdegree and ii) every department treats the same number of customer types, i.e. every node $j \in J$ has the same indegree.*

Finally, let us define a specific class of symmetric networks that are indexed by a single integer value k.

Definition 3 *A k-flexible network has a flexible arc set denoted by $\mathcal{F}_k$. $\mathcal{F}_1 = \{(1, n+1), (2, n+2), ..., (n, 2n)\}$ represents the case of specialized servers. $\mathcal{F}_k$ is constructed on $\mathcal{F}_{k-1}$ as follows: $\mathcal{F}_k = \mathcal{F}_{k-1} \cup \{(1, n+k), (2, n+k+1), (n, n+k+n-1)\}$ where the labeling of the nodes is such that whenever $j > 2n$ we take $j - n$.*

Note that in a k-flexible network k indexes the number of server types that a customer type can be served by (which is equal to the outdegree of the nodes $i \in I$), or equivalently the number of customer types that a server type can treat (which is equal to the indegree of the nodes $j \in J$). In addition, by definition k-flexible networks are always connected.

Throughout, the maximum flow of a network G is denoted by $T_G(\mathcal{F}, \boldsymbol{\lambda}, \mathbf{C})$. For notational compactness, these are replaced by T_G and $T(\mathcal{F})$ whenever possible.

4 Some Properties on Flexibility/Capacity Interactions

One way of improving performance in flexible systems is by adding capacity. For example, one may choose to invest in additional capacity rather than investing in additional flexibility. The close interactions between flexibility and capacity are evident. The following result formalizes this relationship:

Theorem 1 *Consider a symmetric service network with demand vector $\boldsymbol{\lambda}$ and symmetric capacity vector $\mathbf{C}$. If $\sum_{i=1}^{n} \lambda_i > \sum_{i=1}^{n} C_i$ then $T(\mathcal{F}_k, \boldsymbol{\lambda}, \mathbf{C} + \Delta^C) - T(\mathcal{F}_k, \boldsymbol{\lambda}, \mathbf{C}) \geq T(\mathcal{F}_{k-1}, \boldsymbol{\lambda}, \mathbf{C} + \Delta^C) - T(\mathcal{F}_{k-1}, \boldsymbol{\lambda}, \mathbf{C})$ for small enough Δ^C such that one still has $\sum_{i=1}^{n} \lambda_i \geq \sum_{i=1}^{n} C_i + \Delta_i^C$.*

Proof: To show the inequality in the Proposition, we first transform all the networks G with flexibility $\mathcal{F}_k$ into equivalent networks G' with flexibility $\mathcal{F}_{k-1}$. Consider G with demand vector $\boldsymbol{\lambda}$ and capacity vector $\mathbf{C}$. As noted before, $T(\mathcal{F}_k, \boldsymbol{\lambda}, \mathbf{C})$ is the maximum flow of this network. Let $\mathbf{x}^*$ denote the vector of optimum flows in this maximum flow problem. Recall that $\mathcal{F}_k = \mathcal{F}_{k-1} \cup \{(1, n+k), (2, n+k+1), ..., (n, n+k+n-1)\}$. For simplicity denote the arc set $\{(1, n+k), (2, n+k+1), ..., (n, n+k+n-1)\}$ as the set B. Construct a new network G' by taking $\mathcal{F}_k \backslash B$. By definition, this will have the same arc set as $\mathcal{F}_{k-1}$. Let

$$\boldsymbol{\lambda}'_j = \boldsymbol{\lambda}_j + \sum_{\forall i | (i,j) \in \{B\}} x^*_{ij}, \; j = n+1, \ldots, 2n.$$

Note that as constructed $\boldsymbol{\lambda}' \geq \boldsymbol{\lambda}$. Whenever x^*_{ij} for$(i,j) \in B > 0$ in the maximum flow problem of G then in the maximum flow problem of G' with $\mathcal{F}_{k-1}, \boldsymbol{\lambda}, \mathbf{C}$, there will be a slack in capacity at least as large as that in G. By replacing $\boldsymbol{\lambda}$ with $\boldsymbol{\lambda}'$ as constructed, we can ensure filling this slack in capacity. By construction we then have

$$T(\mathcal{F}_k, \boldsymbol{\lambda}, \mathbf{C}) = T(\mathcal{F}_{k-1}, \boldsymbol{\lambda}', \mathbf{C}).$$

For the network with $\mathcal{F}_k, \boldsymbol{\lambda}, \mathbf{C}+\boldsymbol{\Delta}^{\mathbf{C}}$ construct $\boldsymbol{\lambda}''$ in a similar fashion for the corresponding network with flexibility $\mathcal{F}_{k-1}$. By construction we will then have

$$T(\mathcal{F}_k, \boldsymbol{\lambda}, \mathbf{C}+\boldsymbol{\Delta}^{\mathbf{C}}) = T(\mathcal{F}_{k-1}, \boldsymbol{\lambda}'', \mathbf{C}+\boldsymbol{\Delta}^{\mathbf{C}}).$$

Since $\mathbf{C}+\boldsymbol{\Delta}^{\mathbf{C}} \geq \mathbf{C}$, $\mathbf{x}^*$ for the network with $\mathbf{C}+\boldsymbol{\Delta}^{\mathbf{C}}$ will be greater than or equal to the $\mathbf{x}^*$ for the network with $\mathbf{C}$. As a result $\boldsymbol{\lambda}'' \geq \boldsymbol{\lambda}'$. Thus we know that for the maximum flow problem

$$T(\mathcal{F}_{k-1}, \boldsymbol{\lambda}'', \mathbf{C}+\boldsymbol{\Delta}^{\mathbf{C}}) \geq T(\mathcal{F}_{k-1}, \boldsymbol{\lambda}', \mathbf{C}+\boldsymbol{\Delta}^{\mathbf{C}}). \tag{1}$$

Using our construction, the condition we want to prove is equivalent to

$$T(\mathcal{F}_{k-1}, \lambda'', \mathbf{C}+\boldsymbol{\Delta}^{\mathbf{C}}) - T(\mathcal{F}_{k-1}, \lambda', \mathbf{C}) \geq T(\mathcal{F}_{k-1}, \lambda, \mathbf{C}+\boldsymbol{\Delta}^{\mathbf{C}}) - T(\mathcal{F}_{k-1}, \lambda, \mathbf{C}). \tag{2}$$

By Equation (1) we have

$$T(\mathcal{F}_{k-1}, \lambda'', \mathbf{C}+\boldsymbol{\Delta}^{\mathbf{C}}) - T(\mathcal{F}_{k-1}, \lambda', \mathbf{C}) \geq T(\mathcal{F}_{k-1}, \lambda', \mathbf{C}+\boldsymbol{\Delta}^{\mathbf{C}}) - T(\mathcal{F}_{k-1}, \lambda', \mathbf{C}).$$

Showing

$$T(\mathcal{F}_{k-1}, \lambda', \mathbf{C}+\boldsymbol{\Delta}^{\mathbf{C}}) - T(\mathcal{F}_{k-1}, \lambda', \mathbf{C}) \geq T(\mathcal{F}_{k-1}, \lambda, \mathbf{C}+\boldsymbol{\Delta}^{\mathbf{C}}) - T(\mathcal{F}_{k-1}, \lambda, \mathbf{C}) \tag{3}$$

will ensure that the desired condition in (2) holds. It is known that the optimal value of the objective function in the minimum cut problem $(T(\mathcal{F}_i, \lambda, \mathbf{C}))$ is submodular in $(\lambda, -\mathbf{C})$ (Theorem 3.7.1 in [15]) which implies inequality (3). □

Theorem 1 demonstrates that flexibility and capacity are complements up to a certain threshold, parameterized by the demand and capacity vectors. Note that this threshold is equal to $T(\mathcal{F}_n, \boldsymbol{\lambda}, \mathbf{C})$, the throughput of the fully flexible system (recalling that $T(\mathcal{F}_n, \boldsymbol{\lambda}, \mathbf{C}) = \min(\sum_{i=1}^n \lambda_i, \sum_{i=1}^n C_i)$). Beyond the threshold, capacity may act as a substitute to flexibility.

For the case with random demand, this result can only be stated for 'chronically overloaded' systems where total demand always exceeds total capacity for all possible demand realizations. Informally, for heavily utilized systems, which are also those systems where system flexibility is sought most, one will mostly be in the former region with capacity and flexibility acting as complements. For these types of systems, the result suggests that an additional server is more valuable in the system with superior flexibility. Thus flexibility and capacity should be jointly designed. For systems with lower utilization, the result may be reversed, and an additional person can be worth more in a system with less flexibility, reflecting a positive marginal value for capacity.

5 Diminishing Returns to Increased Flexibility

It is well known that the performance of a service system with limited flexibility rapidly approaches that of a system with full flexibility. Numerical examples in a number of articles (e.g. [9], [10]) support this point. This issue will be explored next by investigating structural properties of the maximum flow as a function of $\mathcal{F}$.

The following theorem which establishes the submodularity of the throughput in $\mathcal{F}$ is an adaptation of Theorem 20 in [4]. In order to adapt their result, let X_m $(m = 1, 2, .., n^2)$ be 1 if flexible arc $m \in \mathcal{F}$ and 0 otherwise. Note that X_m is a chain in the sense of partially ordered sets (see [15]). Now let $X = \times_{i=1}^{n^2} X_m$. It is shown in [4] that for sets of substitute arcs (such as those in $\mathcal{F}$), the minimum cost flow is supermodular in X. Since X and $\mathcal{F}$ are equivalent representations of the flexible arc set and the minimum cost flow problem can be converted to a maximum flow problem, we reach the following:

Theorem 2 *: The throughput (maximum flow) of the network, $T_G(\mathcal{F})$ is submodular in $\mathcal{F}$.*

Most manufacturing or service flexibility applications consider the case where the demand, $\boldsymbol{\lambda}$, is a random vector. The following corollary establishes the result in the random demand case.

Corollary 1 *: The expected throughput of the network is $E[T_G(\mathcal{F})]$ is submodular in $\mathcal{F}$.*

Proof: For any realization of the random vector $\boldsymbol{\lambda}$, the throughput is submodular in $\mathcal{F}$ by Theorem 2. Because submodularity is preserved under the expected value operation ([15]), the expected throughput is also submodular in $\mathcal{F}$. □

In most flexibility applications (as in [9]), $\boldsymbol{\lambda}$ is a random vector whereas $\mathbf{C}$ is assumed to be constant. Nevertheless, Corollary 1 directly extends to the case where both $\boldsymbol{\lambda}$ and $\mathbf{C}$ are random.

Submodularity implies that all marginal flexibility improvements are substitutes of each other, thereby suggesting the effectiveness of smart limited flexibility. Submodularity of the expected throughput in terms of the arc set $\mathcal{F}$ reflects one side of the diminishing returns property: marginal additions of flexibility (in terms of arcs) are relatively more beneficial in terms of throughput than combined additions. In general, however, the diminishing returns property is not interpreted in terms of submodularity, but in terms of concavity of the throughput as flexibility increases. Next, we investigate the concavity properties.

Proposition 1 *: If $\mathbf{C} = (C, ..., C)$ and λ_i $(i = 1, 2, .., n)$ are independent and identically distributed random variables, then the expected throughput of the network as a function of $\mathcal{F}_k$ is nondecreasing and concave in k (for $k = 2, ..., n-1$).*

Proof: Recall that $\mathcal{F}_k = \mathcal{F}_{k-1} \cup \{(1, n+k), (2, n+k+1), (n, n+k+n-1)\}$. Since $\mathcal{F}_{k-1} \subset \mathcal{F}_k$, the nondecreasing part of the statement is obvious. Now let us consider the following set: $\mathcal{F}'_k = \mathcal{F}_{k-1} \cup \{(1, n+k+1), (2, n+k+2), ..., (n, n+n+k)\}$. By Theorem 2 and Corollary 1 we obtain:

$$E[T(\mathcal{F}_k \cup \mathcal{F}'_k)] + E[T(\mathcal{F}_k \cap \mathcal{F}'_k)] \leq E[T(\mathcal{F}_k))] + E[T(\mathcal{F}'_k)]. \tag{4}$$

Now note that, by construction : $\mathcal{F}_k \cap \mathcal{F}'_k = \mathcal{F}_{k-1}$ and $\mathcal{F}_k \cup \mathcal{F}'_k = \mathcal{F}_{k+1}$. In addition, $\mathcal{F}'_k$ and $\mathcal{F}_k$ have identical structures by a relabeling of the nodes, which implies that: $E[T(\mathcal{F}'_k)] = E[T(\mathcal{F}_k)]$. Using these equalities, inequality (4) can be expressed as:

$$E[T(\mathcal{F}_{k+1})] + E[T(\mathcal{F}_{k-1})] \leq 2E[T(\mathcal{F}_k))]$$

which is the desired result. □

Proposition 1 proves the concavity property that is observed in the numerical examples of [9] for independent and identically distributed demands. On the other hand, all numerical results seem to indicate that the above concavity holds under even weaker assumptions on the demand distributions. Unfortunately, a general proof under weaker assumptions eludes us. Nevertheless, the next proposition establishes that, in the special case of a 3 by 3 network, concavity of the expected throughput (in terms of the flexibility index) holds for any joint demand distribution. The complete proof can be found in [1].

Proposition 2 *: If $\mathbf{C} = (C, C, C)$ and λ_i $(i = 1, 2, 3)$ are jointly distributed random variables, then the expected throughput of the network has the following diminishing returns property: $E[T(\mathcal{F}_3)] - E[T(\mathcal{F}_2)] \leq E[T(\mathcal{F}_2)] - E[T(\mathcal{F}_1)]$*

6 On the Value of Balance in Flexibility Structures

So far, we have shown basic features of flexibility that would be useful in answering the *how much flexibility* type of question. Next, the theory of majorization [11] is used to explore the type of flexibility, thereby further refining the notion of *smart limited flexibility.*

Definition 4 *For a vector $x \in \mathcal{R}^n$, let $[i]$ denote a permutation of the indices $\{1, 2, ..., n\}$ such that $x_{[1]} \geq x_{[2]} \geq \ldots \geq x_{[n]}$. Then, for x, $y \in \mathcal{R}^n$, x is said to be majorized by y, $x \prec y$, if $\sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} y_{[i]}$ and for all $k = 1, \ldots, n-1$, $\sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]}$.*

Let $i_d(J) \in \mathcal{R}^n$ be the vector of indegrees for $j \in J$, and $o_d(I) \in \mathcal{R}^n$ be the vector of outdegrees for $i \in I$. Recall that the former represents a vector with the number of skills of the servers in each department, and the latter the number of possible routings for each customer class. The following definitions are proposed for the graph $G = (N, A)$.

Definition 5 *If $i_d(J)$ has $i_d(n+1) = i_d(n+2) = \ldots = i_d(2n)$, the service system is said to have balanced skill diversity. Symmetrically, if $o_d(I)$ has $o_d(1) = o_d(2) = \ldots = o_d(n)$, the system is said to have balanced routings. For two networks $G = (N, A)$ and $G'(N, A')$, and skill diversity vectors $i_d(J)$ and $i'_d(J)$ (routing vectors $o_d(I)$ and $o'_d(I)$), whenever $i_d(J) \prec i'_d(J)$ ($o_d(I) \prec o'_d(I)$) the system $G(N, A)$ is said to have more balanced skill diversity (more balanced routings) than $G'(N, A')$.*

Theorem 3 *Consider a network $G(N, A)$ with demand vector $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)$, symmetric capacity vector $\mathbf{C} = (C, ..., C)$, and routing vector $o_d(I)$. Whenever $G(N, A)$ does not have balanced skill diversity, one can find $G'(N', A')$ with demand vector $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)$, symmetric capacity vector $\mathbf{C} = (C, ..., C)$, routing vector $o'_d(I') = o_d(I)$, and $i'_d(J') \prec i_d(J)$ such that $T_{G'} \geq T_G$, as long as the new network G' remains connected.*

Proof: The proof requires some additional definitions and notation and is provided in the Appendix. □

Remark: The final requirement that the new network remains connected is an important one. Without this condition, one could envision a new network with a more balanced skill set, however which consists of multiple shorter chains rather than one long chain as in the original network. When this happens, optimal throughput can be less in the new network, despite the improved skill balance. The following example illustrates this, and provides further support to the earlier claim that longer-fewer chains are better. Consider a symmetric network with $n = 4$. Assume that the flexible arcs are: $\{(1,5),(2,6),(3,7),(3,5),(4,8),(4,5)\}$. Let $\boldsymbol{\lambda} = (0, a, 2a, a)$ and $\mathbf{C} = (a, a, a, a)$. This network can satisfy all the demand, i.e. the maximum flow is 4a. Now construct the following network which is majorized by the original one $\{(1,5),(2,6),(3,7),(3,6),(4,8),(4,5)\}$. Note that this new network consist of two chains rather than just one, and is not connected. Its maximum flow is $3a < 4a$.

Corollary 2 *Consider a network $G(N, A)$ with symmetric demand vector $\boldsymbol{\lambda} = (\lambda, ..., \lambda)$, capacity vector $\mathbf{C} = (C_1, ..., C_n)$, and skill diversity vector $i_d(J)$. Whenever $G(N, A)$ does not have balanced routing, one can find $G'(N', A')$ with symmetric demand vector $\boldsymbol{\lambda} = (\lambda, ..., \lambda)$, capacity vector $\mathbf{C} = (C_1, ..., C_n)$, skill diversity vector $i'_d(J') = i_d(J)$, and $o'_d(I') \prec o_d(I)$ such that $T_{G'} \geq T_G$, as long as the new network G' remains connected.*

The results in Theorem 3 (and Corollary 2) continue to hold for the expected throughput when the demand (respectively the capacity) is random. They formalize similar guidelines suggested in [9], that recommend "equalizing the number of plants (measured in total units of capacity) to which each product in the chain is connected" and "equalizing the number of products (measured in total units of expected demand) to which each plant in the chain is directly connected".

Acknowledgement: The authors would like to thank an anonymous referee for several useful suggestions and the example in Section 6.

References

[1] Akşin, O.Z. and Karaesmen, F. "Designing Flexibility: Characterizing the Value of Cross-Training Practices". *INSEAD, Working Paper*, February 2002.

[2] De Groote, X. "The flexibility of production processes: a general framework". *Management Science*, 40:7 933-945, 1994.

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

[4] Granot F. and Veinott A.F. “Substitutes, Complements and Ripples in Network Flows”. *Mathematics of Operations Research*, 10: 175-186, 1985.

[5] Graves, S. C. and Tomlin B.T. “Process Flexibility in Supply Chains”. *Management Science*, 49: 907-919, 2003.

[6] Gurumurthi, S. and Benjaafar S. “Modeling and Analysis of Flexible Queueing Systems”. *Naval Research Logistics*, 51: 755-782, 2004. .

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

[8] Iravani, S.M., Van Oyen, M.P., and K.T. Sims “Structural Flexibility: A New Perspective on the Design of Manufacturing and Service Operations”. *Management Science*, 51:151-166, 2005.

[9] Jordan, W.C. and Graves, S.C. “Principles on the benefits of manufacturing process flexibility”. *Management Science*, 41:4 577-594, 1995.

[10] Karaesmen F., F. van der Duyn Schouten and L. van Wassenhove, “Dedication vs. Flexibility in Field Service Operations”, *Working Paper, Center for Economic Research, Tilburg University, revised version*, 2003.

[11] Marshall, A. W. and Olkin, I. *Inequalities: Theory of Majorization and Its Applications.* Academic Press, New York, 1979.

[12] Netessine, S., Dobson, G. and Shumsky, R. “Flexible service capacity: optimal investment and the impact of demand correlation”. *Operations Research*, 50:2, 375-389, 2002.

[13] Pinker, E. and Shumsky, R. “Efficiency-quality tradeoff of crosstrained workers”. *Manufacturing and Service Operations Management*, 2:1 , 2000.

[14] Sethi, A.K. and Sethi, S.P. “Flexibility in manufacturing: a survey”. *International Journal of Flexible Manufacturing Systems*, 2:289-328, 1990.

[15] Topkis, D.M. *Supermodularity and Complementarity.* Princeton University Press, Princeton, New Jersey, 1998.

[16] Van Oyen, M.P., Gel, E.G.S., and Hopp, W.J. “Performance opportunity for workforce agility in collaborative and noncollaborative work systems”. *IIE Transactions*, 33:9, 761-777, 2001.

[17] Van Mieghem, J.A. "Investment strategies for flexible resources". *Management Science*, 44:1071-1078, 1998.

A Proofs

The following additional notation and definitions are used in the proof Theorem 3: For a set of nodes I and J, $(I, J) = \{(i, j) : i \in I, j \in J\}$ represents the set of all arcs between I and J. The capacity of an arc (i, j) is given by the capacity function $c(i, j)$. For subsets I and J of N, denote the sum of all capacities on all arcs (I, J) by $c(I, J) = \sum_{i \in I, j \in J} c(i, j)$. Let X be a subset of N. Then X is a cut if it contains the source but not the sink of the network. The cut capacity function is given by $f(X) = c(X, N \setminus X)$. For X being a cut, the minimum of $f(X)$ over all X is known as a minimum cut.

The following lemma is used in the proof of Theorem 3.

Lemma 1 *For the graph $G = (N, A)$ representing the flexible service system, any cut that satisfies at least one of the conditions below cannot be a minimum cut: i) There exists an $i \in X$ with $j \in J$ and $j \notin X$ ii) $f(X) > min(\sum_{i=1}^{n} \lambda_i, \sum_{i=1}^{n} C_i)$.*

Proof: Any cut X that satisfies i) will have $f(X) = \infty$. Since there exists cuts with finite capacity, X cannot be the minimum cut of this network. To show that any cut X that satisfies ii) cannot be a minimum cut, note that both $\sum_{i=1}^{n} \lambda_i$ and $\sum_{i=1}^{n} C_i$ are cuts of this network. □

A.1 Proof of Theorem 3

Take any connected network, characterized by the graph $G(N, A)$. Let NA_G denote the set of cuts of this network, which cannot be eliminated by one of the rules in Lemma 1. This set will be called the set of uneliminated cuts. By Lemma 1, the minimum cut of the network G is a cut in NA_G. Now consider a second network $G'(N, A')$, where an arc (a, b) has been replaced by arc (a, b') and everything else is the same. The nodes $b \in J$ and $b' \in J'$ are chosen such that $i'_d(J') \prec i_d(J)$, and the network remains connected . Then according to Definition 5, the network G' is said to have more balanced skill diversity. It is next shown that G' thus obtained has maximum flow $T_{G'} \geq T_G$.

For any cut $X \in NA_G$ let X_I denote the nodes of X that are in the set I, i.e. $X_I = \{x \in X : x \in I\}$. Recall that $A(X_I)$ denotes the set of adjacent nodes to the nodes in X_I. By Lemma 1, all $y \in A(X_I)$ are also in X, i.e. $y \in X$. $NA_{G'}$ can be obtained from NA_G, by noting that the cuts in NA_G fall into three distinct sets: 1) The set of cuts that do not change as a result of the arc replacement being considered 2) The set of cuts that change, however have the same value as before 3) The set of cuts that are eliminated by Lemma 1 in the new network. In addition, there may be some new cuts.

More precisely, the cuts in NA_G can be grouped as follows. 1) All cuts $X \in NA_G$ such that $a \notin X$ belong to the first group, since the proposed change in the network only impacts $A(a)$. All such cuts will also be in $NA_{G'}$. 2) The cuts in the second group are those that are obtained from a cut $X \in NA_G$ by replacing the node b in the cut by b', i.e. $X' = X - b + b'$, such that for all $x \in X'_{I'}$ $A(X'_{I'}) \in X'$. Thus, these cuts cannot be eliminated by Lemma 1. Note that for this group of cuts, $f(X) = f(X')$ by symmetry of the capacity vector $\mathbf{C}$. 3) Consider a cut X with nodes $x_1 \in X_I$ and $x_2 \in X_I$, and $b \in \{A(x_1) \cap A(x_2)\}$. If for this cut $b' \notin \{A(x_1) \cup A(x_2)\}$, then one will have $\{A'(x_1) \cup A'(x_2)\} \supset \{A(x_1) \cup A(x_2)\}$. Thus any cut $X \in NA_G$ with $x_1 \in X$ and $x_2 \in X$ will have $A'(X'_I) \supset A(X_I)$ in the new network G'. But then all of these cuts will be eliminated by condition i) of Lemma 1. 4) The argument for the third group of cuts shows that there may be some new cuts X' in $NA_{G'}$ with $x_1, x_2 \in X'$ and $A'(X'_I) \in X'$. In other words, these cuts X' contain all the nodes of cuts X that are in group three above, and in addition also contain node b'. Note that for these additional cuts $f(X) + C = f(X')$. Finally observe that if the arc replacement is performed without ensuring that the new network is connected, there may be some cuts X such that $A(X_I) \cap b = \emptyset$, that were eliminated by Lemma 1 in G (i.e. $X \notin NA_G$) but that can become feasible (i.e. $X \in NA_{G'}$) in the new network. One can then no longer guarantee $f(X) \leq f(X')$ for any $\boldsymbol{\lambda}$. Such cuts are avoided by imposing the connectedness condition.

Characterizing the cuts of the new network, using the set of uneliminated cuts for the initial network, one observes that some cuts of G are eliminated in G', while those that are not eliminated preserve the same value $f(X) = f(X')$. All additional cuts that can be added to G' without being eliminated by Lemma 1 are shown to have $f(X') > f(X)$ for some $X \in NA_G, \notin NA_{G'}$. Thus one has that the minimum cut of G' is greater than or equal to the minimum cut of G, which implies that $T_{G'} \geq T_G$. The same argument can be repeated for another arc change that induces more balanced skill sets. Thus any connected network G' can be obtained from a connected network G through a finite number of arc changes, each improving throughput. This proves the result. □

A.2 Proof of Corollary 2

Note that the service networks represented by graphs $G(N, A)$ are fully symmetric in $\boldsymbol{\lambda}$ and $\mathbf{C}$. In other words, a network where the flow is from the sink node towards the source node, and where the capacity vector $\mathbf{C}$ has been replaced by the demand vector $\boldsymbol{\lambda}$ and vice versa, has identical maximum flow with the original network. These latter types of networks can be labeled as the *reversed networks*. Observe, furthermore, that in the reversed network, all results previously shown for skill sets hold, and these are equivalent to results in terms of routings in the original network. Using this equivalence, and noting that the $\boldsymbol{\lambda}$ and $\mathbf{C}$ vectors in the corollary ensure that the reversed network has the same characteristics as the original network (any demand vector, symmetric capacity vector), the result stated in the Corollary follows by the proof for Theorem 3. □