Optimizing Organizational Design in Complex Service Delivery Systems

Yixin Diao
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA

Gargi Dasgupta
IBM India Research Lab
Bangalore 560045, India

***Abstract*— IT service delivery becomes an increasingly challenging business as customers demand improved quality of service while providers are driven to reduce the cost of delivery. In this paper we propose a set of optimization models to recommend organizational design changes in complex service delivery systems. The optimization models take into consideration various design factors such as technology sustainability and management complexity, and provide recommendations to organizational design transformation including both customer focused design and technology focused design. We demonstrate the applicability of the proposed methodology using data from a large IT service delivery environment.**

I. Introduction

Service based economies and business models have gained significant importance over the years. The customers and service providers exchange value through service interactions with the goal of achieving their desired outcomes. Given the focus on the customer's individual value and unique needs, the service providers need to meet a large variety of expectations set by the customers with due diligence. At the same time, they need to continuously evolve better operation methods to minimize the cost of delivery in order to be competitive in the market.

In an effort to improve the quality of services while containing the cost, the service providers rely on a global delivery model, where services are provided to customers from multiple geographically distributed locations either on-shore or off-shore [1]. Despite the advantages such as leveraging qualified local skills and providing resilient and round the clock support, the globally distributed nature of the service support also challenges the service providers to ensure that each service team remains efficient in handling the diverse workload coming from different customers. Although the delivery efficiency can be gained by servicing multiple customers from the same delivery team, this may also raise customer's concerns on the risk of deteriorated services resulted from shared support.

There exists considerable literature using either analytical or simulation based approaches to improve service efficiency and quality in a global delivery environment (e.g., [2], [3]). However, they mainly focus on understanding the interaction between the customer workload and delivery capacity and making staffing related decisions, and are all under the existing service delivery structure on how customers have been grouped together.

In this paper we study how to effectively service customers from the organizational design's perspective. We start from an existing delivery organization and propose a set of optimization models that transform the organization into a more effective design. Specifically, we propose an inner loop model that creates customer focused groups with reduced management complexity subject to technology sustainability, an outer loop model that batches the organizational transformation with minimum impact to the existing delivery structure, and a middle loop model that consolidate the remaining customers into technology focused groups with minimum fragmentation. These models essentially maps a team formation problem where each team comprises of multiple customers grouped together, whose services are delivered together out of one or more geographical locations, considering the technologies that they support, the geographical spread of the work, and the available skills of service agents. Note that the proposed optimization models focus on how to group multiple customers together in order to improve organizational effectiveness and reduce management complexity. As such, the customer request servcie time and the service level agreement are not explicitely considered in these models, but implicitely satisfied when the number of per customer service agents is specified.

There are two main contributions of this paper. First, it studies the general problem of improving organizational management effectiveness and proposes a multi-dimensional optimization framework (so-called "three-loop" design). Such a framework makes it possible to decompose a complex business problem into a set of manageable subproblems with quantifiable metrics and solvable mathematical models. Second, in solving the middle loop optimization problem, we propose an optimization technique with a dual representation of business objectives (both as implicit preference values and as explicit performance indicators) and a cascaded optimization process (for both linear and nonlinear search). We believe such an algorithm provides novel contribution to solving complex optimization problems and have not seen anything similar in existing literature.

The remainder of this paper is organized as follows. Section II overviews the organizational design approaches and the design transformation methodology. Section III presents the

proposed optimization models including the inner loop model, the outer loop model, and the middle loop model that target to different phases of organizational design transformation. Section IV describes the evaluation study to demonstrate the effectiveness and applicability of the proposed models. Section V reviews related work. Our conclusions are contained in Section VI.

II. Organizational Design

A service delivery organization typically services multiple customers belonging to different industries such as banking, telecommunication, and transportation. Providing quality services for all customers while maintaining the cost advantage is one of the biggest challenges for today's service providers. In this section we first overview how service delivery is organized in accordance with one of the following service delivery models: customer focused design and technology focused design [4]. Afterwards, we discuss how to transform the existing delivery structure such that optimal sharing benefits can be realized without risking the deterioration of service quality.

A. Customer Focused Design

In customer focused delivery the service delivery teams are organized according to the customers to be serviced. A single service delivery team is dedicated to support a small set of customers and provides services that cover all technology areas (e.g., network, storage, database) that are interested and contracted with the customers.

The customer focused delivery model tends to build deeper customer knowledge and have higher customer satisfaction. However, it needs to have large enough workload volume and service agents to make it sustainable and cost effective. For example, in order to maintain 24x7 customer support, a minimum of four (and more appropriately, five) service agents are required for each technology area. In addition, due to the workload fluctuation, more service agents are required to handle the peak workload and meet the service level agreements, but during the non-peak periods enough non-time critical workload is also needed to keep the agents fully utilized.

Although sustainability is usually not a concern for large customers, for medium size customers having customer focused delivery implies the need of grouping several customers together. In the latter case, management complexity need to be addressed properly in deciding the proper grouping. The factors that come to consideration typically include the customer industries, the delivery center locations, the workload types and required skills, and the types of service level agreements.

B. Technology Focused Design

In technology focused design the service delivery teams are organized according to the technology areas on which the services are provided. Besides having the advantages of building stronger technical skills, technology focused teams can achieve good multiplexing of work and are most cost effective for both the service providers and the customer.

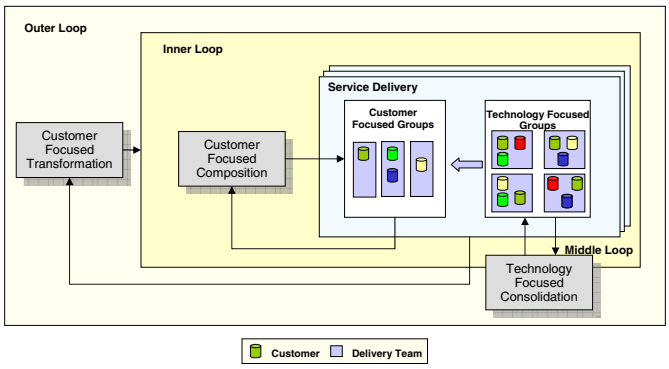

Fig. 1. Illustrative process of organizational design transformation.

However, as the customer's work gets fragmented among multiple delivery teams, integrated service coordination and management become more challenging. This may result in delayed customer response and customer dissatisfaction in complex situations.

C. Design Transformation

The existing delivery organization is generally composed of a mix of service delivery teams with customer focused grouping and technology focused grouping. However, not all of them have been properly formed due to various reasons such as the resource conditions when the contract was first signed and the service scope changes that have occurred over time.

In this paper we propose an optimal organizational design methodology for both the customer focused grouping and the technology focused grouping, as well as a set of optimization models that transform the delivery teams from the existing structure. As illustrated in Figure 1, we refer to these transformations as three optimization loops.

1) Inner Loop Optimization: The inner loop optimization decides the composition of the customer focused groups. The decision of which customers should be in customer focused grouping is typically made in advance based on the size and importance of the customers, as well as the coordination complexity among their contracted technology areas. Given a set of such customer candidates, the objective of the inner loop optimization is form the customer focused groups, such that each group supports its corresponding customers in a sustainable way (with enough workload and agent requirement in all technology areas), while minimizing the management complexity (as indicated by the number of customers, the number of involved delivery centers, and the number of the service agents within each group).

2) Outer Loop Optimization: The outer loop optimization decides how to schedule the customer focused transformation from the existing delivery structure through a few transformation batches. Having a batched (or phased) transformation is needed to reduce the required resources for transformation coordination. Within a batched structure the key consideration in deciding which customer focused groups should go to which

batch is to have minimum impact to the existing delivery teams. For example, if multiple customers within a current delivery team will be affected by the transformation, it would be preferred to make all organizational changes within one batch instead of multiple batches.

3) Middle Loop Optimization: The middle loop optimization decides how to consolidate the remaining customers into technology focused groups. After the formation of the customer focused groups, the remaining service delivery structure often becomes more fragmented in the sense that many small customers are being serviced in the same delivery team. The objective of the middle loop optimization is to minimize such fragmentation, while having less disruption to the normal business operation (e.g., minimizing the needed knowledge transfer for the service agents to work on different customers).

The rationale behind the formulation of these three optimization loops is to decompose the overall organizational design into sub-optimization problems. This is mainly done along the temporal dimension: to decide the customer focused groups first since this is the most important (inner loop), then to schedule the sequence of customer focused transformation (outer loop), and finally to manage the remaining customers into technology focused groups (middle loop). Without such a decomposition, the overall organizational design problem would either be too complex to be solvable, or require more simplifing assumptions will will lead to a less practical solution.

III. Optimization Models

In this section we describe the optimization models for organizational design. We first define the inner loop model that formulates the customer group composition problem as a joint minimization problem and solves it using a greedy algorithm. Next, we discuss the outer loop model that uses the K-means clustering method to partition the customer focused groups defined above into multiple transformation batches. Finally, we describe the middle loop model that organizes the remaining customers into technology focused groups as a linear programming problem. The notation used in this paper is summarized in Table I.

TABLE I

Notation for organizational design models.

$i = 1, 2, \ldots, N$	Set of customers
$j = 1, 2, \ldots, D$	Set of service delivery centers
$k = 1, 2, \ldots, T$	Set of service technology areas
$l = 1, 2, \ldots, G$	Set of customer focused groups
$h = 1, 2, \ldots, H$	Set of transformation batches
$r = 1, 2, \ldots, R$	Set of service delivery teams
x_{ijk}	Number of agents servicing the i-th customer from the j-th delivery center at the k-th technology area
A	Decision matrix with elements a_{il} capturing the composition of customer focused groups
a_{il}	$\begin{cases} 1 & \text{if the } i\text{-th customer belongs to the } l\text{-th group} \\ 0 & \text{otherwise} \end{cases}$
B	Matrix of customer service distribution with elements b_{ij}
b_{ij}	$\begin{cases} 1 & \text{if the } i\text{-th customer is being serviced from the } j\text{-th delivery center} \\ 0 & \text{otherwise} \end{cases}$
C	Matrix of service delivery structure with elements c_{lr}
c_{lr}	$\begin{cases} 1 & \text{if the } l\text{-th customer focused group is being serviced by the } r\text{-th service delivery team} \\ 0 & \text{otherwise} \end{cases}$
Q_h	Set of the h-th transformation batch
μ_h	Cluster center of the h-th batch
n_h	Number of customer focused groups within the h-th batch
p_{ijk}	Assignment preference value for the i-th customer from the j-th delivery center at the k-th technology area
u_{jk}	Unit agent cost at the j-th delivery center with the k-th technical skill
v_{ik}	Total cost for the i-th customer and the k-th technology
w_{jk}	Service capacity for the j-th delivery center and the k-th technology
y_{ijk}	$\begin{cases} 1 & \text{if the } i\text{-th customer is being serviced from the } j\text{-th delivery center with the } k\text{-th technology} \\ 0 & \text{otherwise} \end{cases}$
M	Maximum number of customers that can be grouped together within a customer focused group
S	Minimum number of service agents to form the sustainable customer focused groups

A. Inner Loop Model

Let $i = 1, 2, \ldots, N$ denote the set of customers, $j = 1, 2, \ldots, D$ denote the set of service delivery centers, $k = 1, 2, \ldots, T$ denote the set of service technology areas, and $l = 1, 2, \ldots, G$ denote the set of customer focused groups. Let x_{ijk} denote the number of agents servicing the i-th customer from the j-th delivery center at the k-th technology area.

Let B denote the customer service distribution, where $b_{ij} = 1$ if the i-th customer is being serviced from the j-th delivery center (i.e., $\sum_{k=1}^{T} x_{ijk} > 0$) and $b_{ij} = 0$ if otherwise.

We define A as the decision matrix to capture the composition of customer focused groups, where $a_{il} = 1$ if the i-th customer belongs to the l-th group and 0 if otherwise. The inner loop optimization problem is formulated as follows:

$$\min \max_{l \in G} \sum_{i=1}^{N} a_{il} \tag{1}$$

and

$$\min \max_{l \in G} \sum_{i=1}^{N} \sum_{j=1}^{D} a_{il} b_{ij} \tag{2}$$

and

$$\min \max_{l \in G} \sum_{i=1}^{N} \sum_{j=1}^{D} \sum_{k=1}^{T} a_{il} x_{ijk} \tag{3}$$

s.t.

$$\sum_{i=1}^{N} \sum_{j=1}^{D} a_{il} x_{ijk} \geq S, \forall k, l \tag{4}$$

$$\sum_{i=1}^{N} a_{il} \leq M, \forall l \tag{5}$$

Equation (1), (2), and (3) define the joint minimization objective as to minimize the number of customers within each customer focused group, the number of involved delivery centers within each customer focused group, and the number of service agents within each customer focused group. All these three objectives aim to reduce the management complexity with less coordination in order to provide faster customer response.

Note that while we are forming the customer focused groups, we do not intend to physically relocate the service agents into one delivery center, which is a good approach but often costly to implement and will also have significant disruption to normal service operations. As such, having less involved delivery centers will reduce the coordination work within this virtual group.

Similar to how we use matrix B and Equation (2) to limit the involved delivery centers, we can also include other management complexity limitations such as the customer industries, the workload types and required skills, and the types of the service level agreements. However, in the interest of brevity, we are not showing them as part of the model formulation.

We consider two types of constraints. Equation (4) defines the sustainability constraint where for each technology area k the total number of service agents within each customer focused group l must meet the sustainability criteria S. This is to ensure the cost effectiveness for the formed customer focused group. It also avoids the loss of technical competency by not grouping together enough experts in the same team. In addition, Equation (5) defines another management complexity constraint regarding the total number of customers within each customer focused group l, where M indicates the maximum number of customers that can be grouped together. A smaller number of customer helps the group to build deeper customer knowledge and reduce context switch among different customers.

Since the joint minimization problem defined in Equation (1-5) is NP hard, we propose a greedy algorithm which works as follows:

1) Sort all customers in the descending order based on the total agent size, that is, $\sum_{j=1}^{D}\sum_{k=1}^{T} x_{ijk}$ for customer $i = 1, 2, \ldots, N$.
2) Starting from the largest customer i, check if the sustainability constraint as in Equation (4) can be met for each technology area.
3) If the sustainability criteria can be met for all technology areas, go to step 5); otherwise, find possible grouping customers (starting from 2 and up to M) so that the join of them can meet the sustainability criteria.
4) If more than one possible grouping exists, find the one that involves the least number of service delivery centers, and if still more than one possible grouping exists, find the one with the least number of service agents.
5) Create the customer focused group based on the optimal set of sustainable customers identified above, remove them from the sorted customer list, and go to Step 2) for the next iteration until no more sustainable groups can be identified.

Note that the use of the greedy algorithm can help us incorporate heuristic rules. For example, when grouping different customers, conflicts of interest may arise for the same team servicing competing customers. Also note that in many cases it is not always possible to find the sustainable groups for all customers. If this happens, there are a few possible ways to handle them. First, we can relax the parameter setting that increases the likelihood to find the possible grouping. Second, we can revisit the decision to include them for customer focused grouping and remove them if appropriate. Third, we can add additional agent to make sure they are sustainable, even if it means less cost effective. The final choice will depend on the business needs on a case by case basis.

B. Outer Loop Model

Based on the grouping recommendation from the inner loop model, the objective of the outer loop algorithm is to create the transformation batches with minimum impact to the current service delivery structure.

Let $C = [c_1, c_2, \ldots, c_G]$ denote the service delivery structure for G customer focused groups, where each vector $c_l = [c_{l1}, c_{l2}, \ldots, c_{lR}]'$ indicates how the l-th customer focused group is being serviced by R service delivery teams. That is, $c_{lr} = 1$ if the l-th customer focused group is being serviced by the r-th service delivery team and $c_{lr} = 0$ if otherwise.

We use the K-means clustering algorithm [5] to partition G customer focused groups into H transformation batches $Q = \{Q_1, Q_2, \ldots, Q_H\}$. The algorithm aims at improving the similarity of the customer focused groups within the same transformation batch in terms of the common set of service delivery teams that support these customers. For example, if two customer focused groups happen to be serviced by the same set of service delivery teams, it will be much desired to keep them in the same transformation batch, so that the corresponding delivery teams only have to be affected once.

To run the K-means clustering algorithm, we define the center of cluster Q_h as the mean of all service delivery structure vector within Q_h, that is,

$$\mu_h = \frac{1}{n_h} \sum_{c_l \in Q_h} c_l \tag{6}$$

where n_h is the number of customer focused groups within the h-th batch.

The K-means clustering uses an iterative algorithm that minimizes the sum of distances from each customer focused group to its cluster center, over all clusters. Specifically, we

define the objective function as follows:

$$\min \sum_{h=1}^{H} \sum_{c_l \in Q_h} |c_l - \mu_h|^2 \tag{7}$$

where $|c_l - \mu_h|^2$ indicates the Euclidean distance between the customer focused group vector and its respective cluster center.

C. Middle Loop Model

Following the inner loop and outer loop design on customer focused transformation, the outer loop algorithm consolidates the remaining customers in order to reduce the fragmentation resulted from the transformation.

We achieve customer consolidation by moving and consolidating the current customer workload into fewer delivery centers. (Again we do not intend to physically relocate the service agents as it is costly to implement.) During this consolidation, we also need to consider the knowledge transfer cost that is required to train the agents to work on the customer that they have not worked before.

Let $i = 1, 2, \ldots, N$ denote the set of customers, $j = 1, 2, \ldots, D$ denote the set of service delivery centers, and $k = 1, 2, \ldots, T$ denote the set of service technologies. Note that although we use the same notation i as in the inner loop model to indicate the set of customers, it means a different set of customers. The customers in the inner loop model refer to the customers undergoing the customer focused transformation, and the customers in the middle loop design refer to the remaining customers that will be part of the technology focused groups.

We formulate the middle loop optimization problem as follows:

$$\min \sum_{i=1}^{N} \sum_{j=1}^{D} \sum_{k=1}^{T} p_{ijk} x_{ijk} \tag{8}$$

$s.t.$

$$\sum_{j=1}^{D} u_{jk} x_{ijk} = v_{ik}, \forall i, k \tag{9}$$

$$\sum_{i=1}^{N} x_{ijk} = w_{jk}, \forall j, k \tag{10}$$

$$x_{ijk} \geq 0 \tag{11}$$

where x_{ijk} defines the number of service agents servicing the i-th customer from the j-th delivery center at the k-th technology area, and p_{ijk} defines the corresponding assignment preference value.

Equation (9) defines the equality constraint for the service delivery cost of the customer, where v_{ik} is the total cost for the i-th customer regarding the k-th technology, and u_{jk} is the unit agent cost at the j-th delivery center regarding the k-th technology. Equation (10) defines the equality constraint regarding the service delivery center capacity, where w_{jk} is the service capacity for the j-th delivery center regarding the k-th technology. Overall, these two constraints ensure the customer cost structure and the delivery center agent capacity will not be affected by this consolidation.

Note that we define the middle loop model at the service delivery center level but not the service delivery team level. This is because we separate the consolidation problem into two steps. The first step is to consolidate the customer workload across delivery centers, as defined by the middle loop model in Equation (8-11). The second step is to divide the workload within the delivery center into multiple delivery teams. Since the second step can be done using the K-means clustering algorithm similar to that in the outer loop design, we will not further discuss it in this paper.

The effectiveness of middle loop optimization is centered around how to properly specify the preference values p_{ijk} in order to reflect the business consideration in reducing the fragmentation while having less disruption to their normal service operation (i.e., less knowledge transfer cost).

We start by setting the initial preference values as follows. It uses the current delivery structure (i.e., the current agent assignment x^0_{ijk}) as a reference point.

- If the current assignment $x^0_{ijk} = 0$, set the preference value $p_{ijk} = M$ where M is a very large number (e.g., 9999). This ensures the optimization model will not assign the customer workload into new delivery centers (that the customer has not been serviced before) in order to reduce the fragmentation instead of increasing it.
- If the current assignment $x^0_{ijk} = max\left(x^0_{ijk}\right), \forall j$, set the preference value $p_{ijk} = 0$. This indicates the most favorable delivery center (as the middle loop optimization is defined as a minimization problem). That is, if the majority of the workload (or service agents) are serviced in this delivery center, it is mostly preferred to consolidate other workload into this center as well to reduce both the fragmentation and the knowledge transfer cost.
- Define the preference value for other delivery centers based on their distance to the center delivery center defined above. The delivery center distance can be defined based on the region (or time zone difference) in order to reduce the management complexity. For example, we can define $p_{ijk} = 10$ if from the same region, $p_{ijk} = 20$ if from the same time zone, $p_{ijk} = 30$ if from the neighboring time zone, and so on.

While the above initial preference values implicitly represent the business objectives. We can further develop an iterative algorithm to update the preference values based on two key performance indicators to directly reflect the business needs. Specifically, we define the fragmentation index as

$$\sum_{i=1}^{N} \sum_{j=1}^{D} \sum_{k=1}^{T} y_{ijk} \tag{12}$$

where $y_{ijk} = 1$ if $x_{ijk} > 0$, and $y_{ijk} = 0$ if $x_{ijk} = 0$. This indicates how fragmented the customers are being serviced across the delivery centers.

In addition, we define the knowledge transfer cost as

$$\sum_{i=1}^{N}\sum_{j=1}^{D}\sum_{k=1}^{T}|x_{ijk}-x_{ijk}^{0}| \quad (13)$$

which measures the difference between the new assignment and the current assignment, and a larger difference leads to the need for more knowledge transfer and thus the knowledge transfer cost.

The iterative algorithm is defined as follows:

1) Set the initial preference value using the approach described above based on the current assignment.
2) Solve the linear programming problem [6] described by Equation (8) - (11) to find the optimal solution.
3) Calculate the fragmentation index and the knowledge transfer cost defined in Equation (12) and (13).
4) Update the preference value using the nonlinear search method (e.g., Nelder-mead simplex search [7]) to reduce the fragmentation index and the knowledge transfer cost.

IV. Evaluation

In this section we evaluate the proposed organizational transformation methodology using data from a large IT service delivery environment. We demonstrate how the three optimization models can be built with the available organizational data and provide useful guidance to support the transformation. Note that when necessary the data have been altered to preserve data privacy and simplified for the illustration purpose, though the nature of service operation and organization has been maintained.

A. Inner Loop Evaluation

To build the inner loop model, we need to know for each customer that will be part of the customer focused design how many agents are required for each technology area and which delivery centers they are currently serviced from. This is a comprehensive set of data that can be obtained either from the customer contracts or, more accurately, from the labor claim data where the service agents record how their time is spent.

Table II shows the 18 customer candidates that we will study in the paper and their agent distribution across 7 technology areas (e.g., Application Host Services, Database Management, Intel Platform Support). The agents are defined as FTE, or full time equivalent of work, so that they do not have to be an integer. Although the actual data also include the further breakdown at 11 delivery centers, we will not be able to show them here due to the space limitation.

These data are used to populate x_{ijk} and b_{ij} as the model parameters for the inner loop algorithm. We set $S = 5$ as the minimum number of agents needed to form the sustainable customer focused groups. As seen from Table II, very few customer (in this case, only customer J) is large enough where each of their technology areas requires more than 5 FTEs. The majority of the customers have one or more small technology areas, which becomes the sustainability issues that need to be resolved by grouping multiple customers together.

TABLE II
AGENT DISTRIBUTION FOR 18 CUSTOMER CANDIDATES ACROSS 7 TECHNOLOGY AREAS.

Customer	AHS	DAT	INT	SMO	UNI	SMD	STO	Total
A	16.9	10.5	10.4	2.4	3.9	0.5	17.6	62.2
B	9.3	15.8		3.8		42.1	18.9	89.8
C	12.4	22.0	9.5		11.2	4.0	11.1	70.3
D	2.0	5.0		2.6		9.7	1.0	20.4
E	2.6	0.05		0.03		17.2	9.5	29.3
F	26.1	15.0	1.6	2.6	3.4	8.3	29.8	86.7
G	3.2	16.2	0.7			30.2	5.3	55.6
H	3.5	6.2	6.2		7.3	0.1	10.7	34.1
I	23.5	13.7	0.7	11.8	8.5	5.6	3.7	67.4
J	9.8	5.4	28.0		5.6	5.8	5.6	60.4
K	1.8	18.4	0.5	0.5	10.0	5.7	4.5	41.4
L	3.3	6.2	1.8	1.3	8.1	7.6	11.9	40.1
M	24.2	10.6	1.8	1.2	0.3	5.5	7.5	51.2
N	2.7	1.2	1.9		5.2	0.4	4.9	16.4
O	0.7	1.8	14.2		7.7	13.5	4.7	42.7
P	0.8	14.0				1.4	4.0	20.2
Q	18.2	3.4	9.4	5.8	3.1	19.9	15.0	74.7
R	0.2	8.3	0.01			5.1	4.8	18.4

TABLE III
SUMMARY OF EXISTING SUSTAINABILITY ISSUES AND DELIVERY CENTERS FOR 18 CUSTOMER CANDIDATES.

Customer	Sustainability Issues	Delivery Centers
A	3	3
B	1	5
C	1	8
D	4	4
E	3	6
F	3	6
G	2	4
H	2	5
I	2	7
J	0	5
K	4	7
L	3	5
M	3	6
N	5	5
O	3	4
P	3	5
Q	2	5
R	3	7

Table III summarizes the number of sustainability issues for each customer as well as the number of involved delivery centers.

To run the inner loop model, we set $M = 5$ as the maximum number of customers that can be grouped together within the same customer focused group. The modeling result is shown in Table IV. It includes 6 sustainable customer focused groups with 11 customers selected out of the 18 candidates. Except customer J is sustainable by itself, all other groups have 2 customers grouped together to make them sustainable.

For example, customer B has 1 sustainability issue for the Server Management Other (SMO) technology area. Although there are multiple possible groupings that can help to remove this sustainability issue (as shown in Table V), the combination of customer B and D will result in the best grouping with the minimum number of customers, the involved delivery centers, and the total service agents. All of these contribute to less

TABLE IV

OPTIMAL CUSTOMER FOCUSED GROUPING GENERATED BY THE INNER LOOP MODEL.

Customer Focused Groups	Delivery Centers	Sustainability Issues	Size
J	4	0	60.4
A + F	6	0	148.9
B + D	6	0	110.2
C + P	8	0	90.5
H + Q	6	0	108.8
I + O	7	0	110.1

TABLE V

GROUPING CANDIDATES FOR CUSTOMER B.

Grouping Candidates	Delivery Centers	Sustainability Issues	Size
B + D	6	0	110.1
B + F + Q	8	0	251.2
B + F + C	9	0	246.8
B + F + A	7	0	238.7
B + F + J	8	0	236.9
B + F + O	8	0	219.2
B + F + H	8	0	210.6
B + Q + C	8	0	234.8
B + Q + I	8	0	231.9
B + Q + A	7	0	226.7
B + Q + J	7	0	224.9
B + Q + O	8	0	207.2
B + Q + K	7	0	205.9
B + Q + L	9	0	204.6
B + F + G + O	7	0	274.9
B + F + G + H	7	0	266.2

management complexity.

From Table IV we see there are 7 customers that cannot form the customer focused group under the current parameter setting (i.e., $S = 5$ and $M = 5$). As discussed earlier, there are various ways handle them (e.g., relax the parameter setting, add additional resources) and the final choice will depend on the business needs on a case by case basis.

B. Outer Loop Evaluation

The key data for outer loop optimization is the service delivery structure C, indicating the delivery relationship between the customers and the service delivery teams. This data is usually stored as part of the service delivery organizational document and can be obtained for designing the transformation batches.

Based on this data set, the K-means clustering algorithm partitions the 6 customer focused groups defined in Table IV into 2 transformation batches. The first batch is composed of customer J, customer B + D, and customer H + Q. The second batch is composed of customer A + F, customer C + P, and customer I + O. These batches are determined so that the common set of service delivery teams within each batch is maximized.

TABLE VI

CURRENT AGENT ASSIGNMENT FROM EXISTING DELIVERY STRUCTURE.

	Region 1		Region 2		Region 3	
Customer	1	2	3	4	5	Total
A	3.5	4.5	32.3	12.2	3.9	56.4
B	5.7	1.1	0	3.3	0	10.1
C	12.8	2.4	19.2	0	0	34.4
D	0	0	3.2	13.8	23.9	40.9
E	0	0	22.1	2.1	2.5	26.7
F	0	0	12.2	12.6	15.7	40.5
Total	22	8	89	44	46	209

C. Middle Loop Evaluation

We evaluate the middle loop design by considering a small example composed of 6 customers serviced from 5 delivery centers in 3 regions. For ease of demonstration we only consider one technology area in this paper but the same method is applicable to multiple technologies.

Table VI shows the current agent (FTE) assignment for each customer in each delivery center, where a significant amount of fragmentation is observed (with a fragmentation index of 20). For example, customer A is being serviced from all 5 delivery centers. Various reasons can contribute to this assignment layout including the resource conditions when the contract was first signed and how the service scope has evolved over time.

The objective of middle loop optimization is to reduce this fragmentation while having minimum impact to the normal service operation in terms of the required knowledge transfer. Table VII shows the assignment preference values for each customer with respect to different delivery centers.

For example, the most favorable delivery center for customer A is delivery center 3 (with a preference value of 0). This is because the majority of service agents for customer A is located in delivery center 3. That is, when we want to reduce the fragmentation and move and consolidate the agents to fewer delivery centers, there will be less agent movement and thus smaller knowledge transfer cost.

The second most favorable delivery center for customer A is delivery center 4 (with a preference value of 10), because it is located in the same region as delivery center 3. This gives the advantage that in case we cannot allocate all agents for customer A into one delivery center, they may be split between the two delivery centers within the same region so as to have less management complexity (as compared to being serviced from different regions).

Similarly, delivery center 2 and 1 will be the third and fourth choices since region 1 is neighboring to region 2, and region 3 is farther away so that delivery center 5 will be the last choice.

Table VIII shows the modeling result from the middle loop algorithm. The fragmentation has been reduced for each customer and for each delivery center. For example, customer A is serviced from 3 delivery centers instead of 5, and delivery center 1 is servicing 2 customers instead of 3. The objective

TABLE VII

ASSIGNMENT PREFERENCE VALUES.

	Region 1		Region 2		Region 3
Customer	1	2	3	4	5
A	60	50	0	10	100
B	0	50	100	90	100
C	10	50	0	100	100
D	9999	9999	100	10	0
E	9999	9999	0	10	100
F	9999	9999	50	50	100

TABLE VIII

RECOMMENDED AGENT ASSIGNMENT FOR TECHNOLOGY FOCUSED GROUPS.

	Region 1		Region 2		Region 3	
Customer	1	2	3	4	5	Total
A	0	4.9	46.6	4.9	0	56.4
B	10.1	0	0	0	0	10.1
C	11.9	3.1	19.4	0	0	34.4
D	0	0	0	0	40.9	40.9
E	0	0	23.0	3.7	0	26.7
F	0	0	0	35.4	5.1	40.5
Total	22	8	89	44	46	209

function value $\sum_{i=1}^{N}\sum_{j=1}^{D}\sum_{k=1}^{T} p_{ijk}x_{ijk}$ in Equation (8) has also dropped from 5086 in the current assignment to 2885 in the recommended assignment, indicating the overall level of fragmentation that has been reduced. Specifically, the fragmentation index has dropped from 20 to 12.

V. RELATED WORK

There is various research aimed to improve the quality and performance of service delivery systems. [8] and [9] solves the change scheduling problem by using a business-driven approach that evaluates change schedules in terms of the financial loss. [10] proposes a change scheduling optimization model that can be solved using standard mixed integer programming techniques. [11] develops a decision support tool to evaluate the impact from business strategies (e.g., different policies for critical incident prioritization). [12] studies request dispatching priorities to meet service attainment targets from multiple customers. However, most of them focus on managing the customer service workload but not on optimizing the delivery center organization.

In another class of literature called workforce management, [13] considers the problem of long term workforce planning with general nonstationary arrival and service time processes. [14] studies dynamic staffing in a call center environment where the objective is high service level attainment. [15] proposes a method for determining the optimal numbers of permanent versus temporary staff and the threshold value at which temporary staff should be called upon, considering conflicting objectives of meeting service level constraints and minimizing costs. [16] studies asymptotically optimal solutions for large scale service systems with multiple customer classes. Although these papers include aspects of managing the workforce in a dynamic service delivery environment, all of them assume the current organization model (i.e., which customers are serviced by which service teams) but not on aim to improve its design and composition.

With respect to organizational design models, [17] studies the shared service model and shows that shared services not only reduce management cost but also improve service quality. On the other hand, [4] and [18] indicate that shared services, if not designed efficiently, can also affect customer satisfaction. There is also work on organizational design principles underlying an effective service delivery model [19] [20] as well as resource hiring and cross-training in such models [21]. However, none of them have formulated the detailed optimization model on how to optimize and transform the organizational design.

VI. CONCLUSIONS AND FUTURE WORK

IT service delivery becomes an increasingly challenging business as customers demand improved quality of service while providers are driven to reduce the cost of delivery. In this paper we have studied a set of optimization models to provide recommended organizational design in complex service delivery systems. This includes both the customer focused design with the advantage to build deeper customer knowledge and have higher customer satisfaction and the technology focused design for building stronger technical skills and being more cost effective through resource sharing.

We proposed an inner loop model that creates customer focused groups with reduced management complexity subject to technology sustainability, an outer loop model that batches the organizational transformation with minimum impact to the existing delivery structure, and a middle loop model that consolidate the remaining customers into technology focused groups with minimum fragmentation. We also demonstrated the applicability of the proposed methodology using data from a large IT services delivery environment.

While the initial results are encouraging, there are several areas for further improvement. First, we would like to further evaluate the method effectivenesss by using different examples and by comparing it with other approaches (e.g., exhaustive search or genetic algorithm for inner loop optimization). This will help to give us a better understanding of the optimality and scalability of the proposed inner loop greedy algorithm. Second, we are working to initiate pilot studies that can help us to further validate the accuracy and usability of the model, and identify and address other practical concerns in organizational design. Third, we would like to develop a user friendly tool so that the model can be run easily by the delivery center management team when changes happen such as new customers come onboard. Finally, we plan to extend the model to consider virtual organizational design for the technology focused design where the delivery team can be composed with resources from different physical delivery locations.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Jason Gast, James Tyler, Tom Lubeck, Rodney Wallace, and Larisa Shwartz, all employed by IBM, for helpful and constructive discussions that helped us improve the quality of the model.

REFERENCES

[1] A. Bose, A. Heching, and S. Sahu, "A framework for model-based continuous improvement of global IT service delivery operations," in *Proceedings of the IEEE International Conference on Services Computing*, 2008, pp. 197–204.

[2] Y. Diao and A. Heching, "Staffing optimization in complex service delivery systems," in *Proceedings of 7th International Conference on Network and Service Management, Paris, France*, 2011.

[3] Z. Feldman and A. Mandelbaum, "Using simulation based stochastic approximation to optimize staffing of systems with skills based routing," in *Proceedings of the 2010 Winter Simulation Conference*, J. M.-T. j. H. B. Johansson, S. Jain and e. E. Yucesan, Eds. Baltimore, MD: The Society for Computer Simulation International, 2010, pp. 3307–3317.

[4] S. Agarwal, R. Sindhgatta, and G. B. Dasgupta, "Does one-size-fit-all suffice for service delivery clients," in *Proceedings of International Conference of Service Oriented Computing*, 2013, pp. 177–191.

[5] B. Mirkin, *Mathematical Classification and Clustering*. Springer, 1996.

[6] S. I. Gass, *Linear Programming: Methods and Applications*. Courier Dover Publications, 2003.

[7] C. T. Kelley, *Iterative Methods for Optimization*. SIAM, 1999.

[8] R. Reboucas, J. Sauve, A. Moura, C. Bartolini, and D. Trastour, "A decision support tool to optimize scheduling of IT changes," in *Proceedings of IFIP/IEEE International Symposium on Integrated Network Management, Munich, Germany*, 2007.

[9] J. Sauve, R. Reboucas, A. Moura, C. Bartolini, A. Boulmakoul, and D. Trastour, "Business-driven support for change management: Planning and scheduling of changes," in *Proceedings of IFIP/IEEE International Workshop on Distributed Systems: Operations and Management, Dublin, Ireland*, 2006.

[10] L. Zia, Y. Diao, D. Rosu, C. Ward, and K. Bhattacharya, "Optimizing change request scheduling in IT service management," in *Proceedings of IEEE International Conference on Services Computing*, 2008.

[11] C. Bartolini, C. Stefanelli, and M. Tortonesi, "Business-impact analysis and simulation of critical incidents in IT service management," in *Proceedings of IFIP/IEEE International Symposium on Integrated Network Management*, 2009.

[12] Y. Diao and A. Heching, "Closed loop performance management for service delivery systems," in *Proceedings of IFIP/IEEE Network Operations and Management Symposium*, 2012.

[13] O. B. Jennnings, A. Mandelbaum, W. A. Massey, and W. Whitt, "Server staffing to meet time-varying demand," *Management Science*, vol. 42, pp. 1383–1394, 1996.

[14] W. Whitt, "Dynamic staffing in a telephone call center aiming to immediately answer all calls," *Operations Research Letters*, vol. 24, pp. 205–212, 1999.

[15] A. Bhandari, A. Scheller-Wolf, and M. Harchol-Balter, "An exact and efficient algorithm for the constrained dynamic operator staffing problem for call centers," *Management Science*, vol. 54, pp. 339–353, 2008.

[16] M. Armony, I. Gurvich, and A. Mandelbaum, "Service level differentiation in call centers with fully flexible servers," *Managment Science*, vol. 54, pp. 279–294, 2008.

[17] T. H. Group, "Shared services and outsourcing network," in *Global service center benchmark study*, 2009.

[18] G. B. Dasgupta, R. Sindhgatta, and S. Agarwal, "Behavioral analysis of service delivery models," in *Proceedings of International Conference of Service Oriented Computing*, 2013.

[19] S. Agarwal, V. K. Reddy, B. Sengupta, S. Bagheri, and K. Ratakonda, "Organizing shared delivery systems," in *Proceedings of International Conference on Services in Emerging Markets*, 2011.

[20] S. Alter, "Service system fundamentals: Work system, value chain, and life cycle," *IBM Systems Journal*, vol. 47, pp. 71–85, 2008.

[21] D. Subramanian and L. An, "Optimal resource action planning analytics for services delivery using hiring, contracting and cross-training of various skills," in *Proceedings of International Conference of Services Computing*, 2008.