Staffing Optimization in Complex Service Delivery Systems

Yixin Diao and Aliza Heching
IBM Thomas J. Watson Research Center
P.O. Box 704, Yorktown Heights, NY 10598, USA
Email: {diao|ahechi}@us.ibm.com

Abstract— **Enterprises and service providers are increasingly challenged with improving the quality of service delivery while containing the cost. However, it is often difficult to effectively manage the conflicting needs associated with dynamic customer workload, strict service level constraints, and service personnel with diverse skill sets. In this paper we propose an optimization model to provide recommended staffing levels in a complex service delivery system. The optimization model minimizes the total staffing related variable cost while considering the contractual service level constraints, the skills required to respond to different types of service requests, and the shift schedules that the service agents must follow. We describe how model-based decision making can be conducted using a combination of optimization and discrete event simulation techniques. We demonstrate the applicability of the proposed approach in a large IT services delivery environment.**

I. INTRODUCTION

In recent years, the IT services industry has faced continual pressure to improve the quality of its services while simultaneously reducing the cost of services delivery. These apparent conflicting objectives of improving quality and reducing cost have lead the industry to explore innovative methods for managing its business. Common metrics for measuring quality of service, including equipment availability, time to resolve incidents, and mean time between failures, are measured on a regular basis as part of the standard management and operating process. In an effort to improve the quality of their services, service providers are adopting a more continual and measurable focus on their internal processes, the skills of their people, and the organizational structure. On the other hand, the inherently labor intensive nature of the IT service industry results in complicated tradeoffs in order to ensure customer requests are satisfied within the contractually specified service quality targets, as well as staffing decisions regarding agent skills, cross-training, temporary labor hiring, and the like. In this paper we propose an optimization framework to support staffing decision making in complex service delivery systems, which takes into consideration the relationships among the customer workload, the contractual service level constraints, the agent skills for service team organization, and the shift schedules to satisfy both service coverage requirement and the local regulatory requirements.

We consider a service delivery system in the context of global IT services delivery. Global delivery refers to a model for delivering IT services where the services provider may provide services from either on-shore or off-shore locations to customers who may be globally located. Figure 1 describes the process whereby customers contract for services with the services provider. A customer may have one or more IT needs such as managing a network, supporting a database, requiring backup and restore services, etc. The customer reviews the menu of IT services offered by the services provider and contracts for one or more services. The infrastructure supporting the services (e.g., servers, networks, application, business processes) may be owned by the customers and located on customer sites; alternatively, it may be owned and located on provider sites on behalf of customers. The service delivery provider has delivery centers that are globally located. As service requests arrive from the customer, the requests are assigned to a service delivery location; agents in the service delivery locations are responsible for handling and responding to the service requests. Although the agents in these service delivery locations respond to the service requests, they do not directly interact with the end customers.

Global service delivery offers both advantages and challenges. Globally located agents allow customers to leverage qualified local skills in each location where the service provider maintains a presence, and allow the service provider to offer round-the-clock support. In the face of natural disasters, a globally distributed agent base improves a provider's resiliency by having distributed support teams and data centers. On the other hand, global delivery challenges a provider to ensure that processes are consistent across the services delivery environment and that high quality service is uniformly provided by all agent teams. The globally distributed nature of the teams also challenges the service provider to ensure that each team remains efficient. While efficiency may be gained by servicing multiple customers from the same delivery center, the service provider is challenged to understand the interaction between the arrival rates of the different requests for service, the team-specific service rates for the different types of service requests, the available skills from the service agents, and the different service level target requirements. This understanding is essential to determine the required staffing levels and shift schedules that minimize the overall labor cost of delivery. We focus on this latter challenge and use an optimization model to support staffing decision making.

We are addressing the staffing decision problem for the system where different requests need to be serviced by service

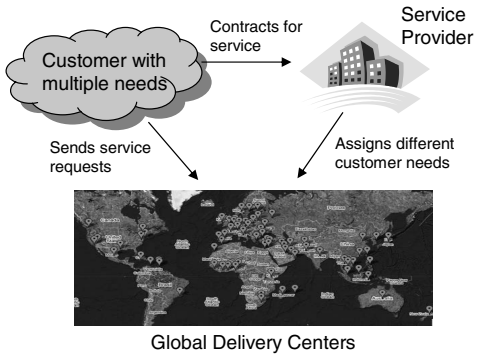


Fig. 1. IT Global Service Delivery.

agents with specific skills. This lies in the staffing optimization area so-called Skills Based Routing ("SBR"). The SBR problem is known to be analytically complex with limited theoretical results. [1] and [2] provide detailed surveys of the analytical approaches that have been undertaken. Common approaches are to simplify the topology of the network or simplify the routing schemes. However, these are not desirable solutions in a service delivery environment where both the network and the routing schemes are complex and the service providers are seeking practical solutions rather than conceptual guidance.

An alternative solution methodology that has applied to the SBR problem is a simulation-based approach. Simulation derives suggested solutions after considering the complexities of the real world system such as the nonstationarities in the arrival rates and the interactions between decisions made in different periods. [3] considers a multi-period problem of determining optimal staffing levels while meeting service level requirements. They solve a sample average approximation of the problem using a simulation based analytic center cutting plane method and assuming that the service level functions are pseudoconcave. [4] extends this approach by applying it to large problem instances and developing heuristic methods to handle the numerical challenges that arise. [5] uses stochastic approximation to determine optimal staffing levels, assuming that the service level functions are convex in the staffing levels. [6] considers a two stage approach for determining optimal staffing levels in a call center environment. In the first stage they solve for the staffing levels by using the per period attainment as an approximation for the true service level attainment. In the second stage, the simulation is used to evaluate true system performance and service level attainment.

While vast literature exists for solving the SBR problem in the context of manufacturing systems and call centers, very few research has been conducted in service delivery systems due to the complexity of customer workload. There are various work types involved in a service delivery system such as incidents, changes, maintenance, project work, etc., and each of them has distinct characteristics and service level requirements. [7] solves the change scheduling problem by using a business-driven approach that evaluates change schedules in terms of the financial loss. [8] proposes a change scheduling optimization model that can be solved using standard mixed integer programming techniques. [9] develops a decision support tool to evaluate the impact from business strategies (e.g., different policies for critical incident prioritization). However, none of them has addressed the optimal staffing problem.

In this paper we propose a staffing optimization approach that determines minimum staffing requirements while meeting contractual service quality commitments. The approach leverages the use of discrete event simulation to capture the complex relationships within a service delivery system. It is worthy to note that the focus of this paper is not on proposing new simulation or routing algorithms, but on how the existing methodology can be successfully applied (and deployed in a large scale) in a global services delivery environment. The proposed approach requires a reasonable set of operational and demographic data which lends its applicability in real service delivery environments. Afterwards, we use scatter search combined with tabu search to solve the optimization model. This propose optimization framework is implemented and deployed at a large services delivery provider with worldwide delivery locations and international customers.

The remainder of this paper is organized as follows. Section II discusses the background for service delivery systems and the challenges for optimization. Section III presents the proposed optimization model and solution approaches. Section IV describes the results of experimental studies that explore the effectiveness and value of the proposed approach. Our conclusions are contained in Section V.

II. Service Delivery Systems

Service delivery involves customers contracting with a services provider on a menu of IT services such as security patch management, network management, and data backup and restore management. The customer contract specifies the scope of services (e.g., number of servers, number of users), the locations from which services will be provided (customer site, provider located), and the measures of quality of service (i.e., service level targets). The service delivery provider responds by assigning each contracted service to a delivery location and maintains a team of service agents to respond to customers' service requests. These teams of agents are typically, though not necessarily, shared across multiple customers. The agents typically are differentiated with respect to the depth and breadth of the skills they have, where the breadth of skill refers to the range of IT areas in which the agent has knowledge and the depth of skills refers to the level of knowledge mastered by the agent in each of these IT areas. Agents are grouped into teams where all agents in a team have common breadth and depth of skills.

We now provide a more detailed description of the workload management process after customers contract for service and once customer requests begin to arrive to the service delivery provider. Figure 2 illustrates the process and operational flow. The arrived service requests are routed to a *service functional unit* at a global delivery location. A service functional unit is

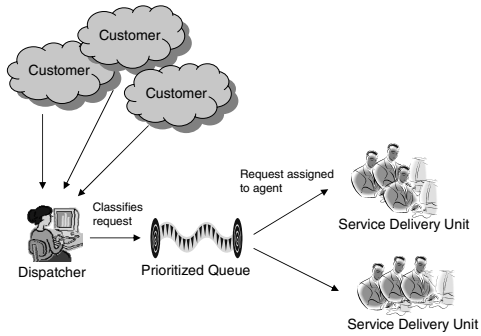


Fig. 2. Illustrative process and operation flow of a service functional unit.

a group of service delivery units with similar IT knowledge, i.e., common breadth of skills. The agents in a service delivery unit can be colocated or working as a virtual team. In the virtual team, all agents are not physically colocated but they function as a team and their performance is jointly measured. A dispatcher is assigned to the service functional unit whose role is to review all incoming requests assigned and determine the priority and the complexity of the requests. Priority of the requests will determine the order in which the requests will be serviced; complexity assigned to the request will determine which service delivery unit is capable of handling the request. The dispatcher may maintain separate queues for each service delivery unit or may maintain a single queue for all service delivery units. The benefit of maintaining a single queue is that some of the requests may be serviced by more than one service delivery unit (e.g., tasks that require low skills may also be assigned to the high skill service delivery unit, following specific rules).

In addition to the scope of service requests that the provider will service for the customer, customer contracts specify service levels associated with each of these service requests. Service levels are a measure of quality of service delivery. Although many types of service level agreements exist, the most common service level agreements specify the following main terms regarding response to a service request: (i) scope of agreement, (ii) target time, (iii) percentage attainment, and (iv) time frame over which service will be measured. For example, a service level agreement may state that 95% (percentage attainment) of all severity 1 tickets (scope) that are opened over each one month period (time frame) must be resolved within 3 hours (target time). One will typically find a large number of service level agreements associated with each customer contract.

Customer service requests can be broadly classified into two types: primary requests and project requests. Primary requests are characterized by relatively short service time (typically, minutes or hours) and short target time (typically, hours or days), and in most cases require a single agent to complete the request. Examples of primary requests include problem tickets, change requests, and maintenance work. Project requests are characterized by requests that are composed of a sequence of tasks and that may require the coordination of a number of different delivery units where different units are responsible for different tasks in the overall project request. There may be dependency relationships between the different tasks. Tasks within a project often take weeks or months to complete. In many cases, the agents who service the project workload are different from those who service the primary workload. This is due to the different skills required. In other cases, the agents are separated into different service delivery units due to the differences in cadence and arrival processes for these two types of workload and the ease in management that is introduced by separating these two types of workload.

In this paper, we focus our study on service delivery units responding to the customer primary requests. Customer requests arriving to the system have a number of attributes including the associated customer, priority (severity), and required skills. The combination of customer, priority, and request type determine the target response time and associated percentage attainment. With this information, all arriving service requests are assigned to a priority class and we assume that the arrival rates of workload to each of the different priority classes is independent.

There are a number of complexities modeling a service delivery system. For example, the number of classes of requests arriving to the system is large, since the requests are differentiated by the different attributes, and the request arrival rates are non-stationary, varying over the hours of the day and days of the week. Second, the agents form different service delivery units have different breadth and depth of skill and are working on different shift hours. Third, the service target time can be measured either against calendar hours or business hours; in the latter case, a business calendar is required. It is due to these modeling complexities as well as the inherent stochastic nature of the problem that we choose a simulation-based modeling and optimization framework to determine optimal staffing levels.

III. Optimization Model

In this section we describe the optimization model for staffing decisions in a global service delivery system. The objective is to minimize the total staffing related variable cost while considering the contractual service level constraints, the skills required to respond to different types of service requests, and the shift schedules that the service agents need to follow. We first define the basic elements and assumptions in our model. Next, we specify the objective function and the constraints, and introduce the simulation model that we use to characterize the delivery system operation. Finally, we discuss the solution techniques. The notation used in this paper is summarized in Table I.

A. Definitions and Assumptions

Let $i = 1, \ldots, I$ denote the set of customers serviced in a service functional unit. We define the optimization model at the level of service functional units, where one model is used for each service functional unit. Although a service request may be routed between multiple service functional units (e.g.,

TABLE I

NOTATION FOR STAFFING OPTIMIZATION.

$i = 1, \ldots, I$	Set of customers
$j = 1, \ldots, J$	Set of shift schedules
$k = 1, \ldots, K$	Set of service delivery units
$c = 1, \ldots, C$	Set of request complexities
$p = 1, \ldots, P$	Set of request priorities
$t = 1, \ldots, T$	Set of intervals in the workload horizon
$l = 1, \ldots, L$	Set of intervals in the shift schedule horizon
$r = 1, \ldots, R$	Set of staffing equality constraints
x_{jk}	Number of agents assigned to shift schedule j in service delivery unit k
$\overline{x}_k$	Upper bound on the number of agents assigned to service delivery unit k
$\underline{x}_k$	Lower bound on the number of agents assigned to service delivery unit k
$\overline{x}_j$	Upper bound on the number of agents assigned to shift schedule j
$\underline{x}_j$	Lower bound on the number of agents assigned to shift schedule j
c_k	Cost of agent in service delivery unit k
b_r	Staffing equality constant for the r-th constraint
a_{kr}	Staffing equality parameter for the r-th constraint at delivery unit k
v^t_{ipc}	Average volume of workload in period t for customer i with priority p and complexity c
s_{ipc}	Average service time for customer i with priority p and complexity c
y_{jl}	$\begin{cases} 1 & \text{if shift } j \text{ is staffed in period } l \\ 0 & \text{otherwise} \end{cases}$
α_{ip}	SLA attainment target for customer i of priority p
w_{ip}	SLA target time for customer i of priority p
m_{ip}	SLA measurement period for customer i of priority p

due to misroutes), the percentage of misrouted requests is extremely small for a well organized service delivery organization. Therefore, we assume no dependencies among the workload or agents in different service functional units and model each of the service functional units independently (i.e., as an open queueing network).

Let $k = 1, \ldots, K$ denote the set of service delivery units in which the service agents are organized within the service functional unit. All agents within the same service delivery unit have the same depth and breadth of skills (e.g., technical skills, customer environment familiarity) and can respond to the arrived service request equally. Let $j = 1, \ldots, J$ be the set of allowable shift schedules to which agents could potentially be assigned. These shift schedules are ensured to provide adequate customer service coverage (e.g., 24 by 7) and to satisfy the regulatory requirements (e.g., total number of hours worked per week or consecutive working hours per day).

Arriving service requests are classified by customer i, as well as by complexity $c = 1, \ldots, C$ based upon the skills required to respond to these requests. The request complexity can be defined to have a one to one mapping to the service delivery unit defined above; alternatively, a mapping table can be used for more complicated mapping. The service requests are further distinguished by their priorities $p = 1, \ldots, P$. We use priority to characterize the severity and urgency of the service requests. The priority level specifies the order of which the arrived request will be processed by the service agent.

Note that given the complexity of the service delivery system a sizable number of parameters are used in the model formulation. However, we design the model such that all required parameters can be commonly measured. This is evidenced in our deployment effort encompassing about 300 service functional units.

B. Objective Function and Constraints

The staffing optimization problem can be formulated in two different ways. The first formulation includes both the staffing cost and the SLA violation cost in the objective function and the goal is to minimize the sum of these two costs. This allows the service delivery provider to trade off between the cost of hiring additional staff versus incurring additional service quality penalties due to lack of sufficient staffing. The second formulation only includes the staffing cost in the objective function but models the service level agreement as a constraint that must be satisfied.

Although mathematically feasible, the first approach is less likely to be adopted by the service provider from the business point of view. Quality of service and service level attainment are metrics that both service delivery providers and customers monitor on a continual basis. Trading off staffing cost against service level attainment does not consider the cost of good will and may very likely lead to customer dissatisfaction. We therefore adopt the latter formulation and state the optimization problem as follows:

$$\min \sum_{j=1}^{J} \sum_{k=1}^{K} c_k x_{jk} \tag{1}$$

$s.t.$

$$f_{ip}(v^t_{ipc}, s_{ipc}, y_{jl}, x_{jk}, w_{ip}, m_{ip}) \leq \alpha_{ip} \tag{2}$$

$$\underline{x}_k \leq \sum_{j=1}^{J} x_{jk} \leq \overline{x}_k \tag{3}$$

$$\underline{x}_j \leq \sum_{k=1}^{K} x_{jk} \leq \overline{x}_j \tag{4}$$

$$\sum_{j=1}^{J} \sum_{k=1}^{K} a_{kr} x_{jk} = b_r \tag{5}$$

$$x_{jk} \geq 0 \tag{6}$$

Equation (1) defines the staffing cost to be minimized where x_{jk} denotes the number of service agents organized in the k-th service delivery unit and assigned to the j-th shift. We also define the cost variables c_k as the unit cost per service agent within the k-th service delivery unit (noting that the delivery units are organized based on the depth and breadth of the skills and highly skilled agents generally demand higher cost). The total staffing cost is summed over agents from all delivery units at all shifts.

We consider the following two types of constraints: service level constraints and staffing coverage constraints. *Service*

level constraints, as defined in Equation (2), represent the service level objectives that must be satisfied. In a service delivery environment, the service level objectives typically takes on a form such as "no more than 5% of priority 1 incidents reported each month can be resolved in more than 2 calendar hours." We use α_{ip} to denote the attainment target associated with the class of service requests from customer i with priority p, w_{ip} to define the SLA target time, and m_{ip} to represent the measurement period. Due to the complexity of service delivery operation, we use discrete event simulation $f(\cdot)$to compute the service attainment level (or, more precisely, the service violation level) from a number of factors.

We define v_{ipc}^{t} as the volume of service requests arriving from customer i with priority p and complexity c during the time period t. The variability in the arriving workload is stochastic in nature over short periods of time but exhibits a repeating weekly pattern. We model the arrival of workload as a non-homogeneous Poisson process. That is, we assume the arrival rate follows a stationary Poisson arrival process within each of one hour time periods for $t = 1, \ldots, T$ ($T = 168$) hours of the week.

We note that some other temporal patterns also exhibit in the data (e.g., due to quarterly or end of year changes). However, we choose not to consider these for the following reasons: First, due to the dynamic nature of the service delivery environment, it is typically difficult to obtain long periods of historical data that are stable enough to derive the seasonal patterns. Second, although seasonal workload patterns do exist, the service delivery units are typically able to use alternative means (other than changes in staffing) to manage the end of quarter / end of month peaks in workload. For example, certain non-demanding workload types (e.g., documentation update, knowledge transfer) can be scheduled during non-peak periods. Third, scheduled overtime can be used to meet some excess demand during periods of high demand, though overtime is not relied upon too extensively due to either regulatory constraints or the already long shifts worked by agents. Finally, the agents' vacations are typically scheduled in consideration of these known seasonal workload patterns. Thus, for example, more agents take vacation in August (Europe) and December when volume of workload decreases for many of the service delivery units, coincident with the time when most of the customers are on vacation. In summary, the weekly workload pattern has been shown to be sufficient to satisfy our practical needs in the context of staffing optimization.

In addition to the workload volume, the simulation model also takes the request service time s_{ipc} for customer i with priority p and complexity c. Similar to the findings of [10] and supported by the theoretical work of [11], based on data collected from many service delivery units we find that the distribution of the service times is well modeled by a lognormal distribution. Finally, we use y_{jl} to denote the shift working hours for schedule j at period l; $y_{jl} = 1$ if shift schedule j is staffed in period l and 0 otherwise (where $l = 1, \ldots, L$ and L defines the periodicity after which the schedule repeats itself).

Staffing coverage constraints, as defined in Equation (3-5), represent the restrictions on the staffing assignment. Equation (3) places restrictions on the number of agents within each service delivery unit. For example, there may be a limited number of high skilled agents available so that the maximum size of the "high skill" service delivery unit is limited. We use $\overline{x}_k$ to state the upper bound of the number of agents in delivery unit k; similarly, a lower bound $\underline{x}_k$ is defined.

In some cases, there are constraints on the number of agents who must work in a given shift. These may be "physical constraints" due to the configuration of the delivery environment. For example, agents may be required to monitor consoles and due to the physical layout of the delivery environment there is a minimum number of agents who must be available to monitor the consoles irrespective of the volume of workload. In other examples, customer contracts may specify a minimum number of agents who must be available during certain shifts. Equation (4) are used to capture these constraints where $\overline{x}_j$ and $\underline{x}_j$ denote the upper bound and lower bound of the number of agents in shift j.

Finally, there are cases where the number of agents from certain delivery units is fixed or a mix of them is fixed, these are captured in Equation (5) as equality constraints where b_r defines the equality constant for the r-th constraint and a_{kr} defines the corresponding coefficient for the k-th delivery unit.

C. Solution Approaches

The staffing optimization problem described in Equation (1-6) defines a feasible region (or a search space) within which one can use analytical techniques to find the optimal staffing levels. Given the complexity of service delivery, we rely on the simulation model to accurately represent the delivery system dynamics and compute the service attainment in order to determine feasibility of proposed solutions.

Considering the implementation practicability, we adopt the intelligent search procedures inherent in the *Scatter Search* combined with *Tabu Search* metaheuristics (see, e.g., [12] [13]). Scatter search originated from strategies for creating composite decision rules and surrogate constraints. It generates a reference set of trial solutions, and improve the solutions by joining solutions based on generalized path constructions in Euclidean space. To enhance the convergence of scatter search, tabu search is applied to recall the performance of proposed solutions that have been evaluated and guide the search process. It ensures that solutions that have already been evaluated will not be reevaluated, guides intensification or diversification of the search, and leads the search away from a locally optimal solution.

The optimizer based on the above approaches makes no mathematical assumptions for the functions within the feasible region. This makes it particularly useful for solving the optimization problem with embedded simulation models, where the solutions suggested by the optimizer are evaluated by the simulation model for performance, and the output of the

TABLE II
SCENARIO OF SERVICE REQUESTS.

	Alerts	Problems	Ad Hoc
Average Weekly Volume	6784	403	940
Average Service Time (min)	5.47	12.40	20.30
Stdev Service Time	3.46	11.77	23.48
Target Response Time (min)	10	30	60
Service Attainment Target	99%	95%	95%

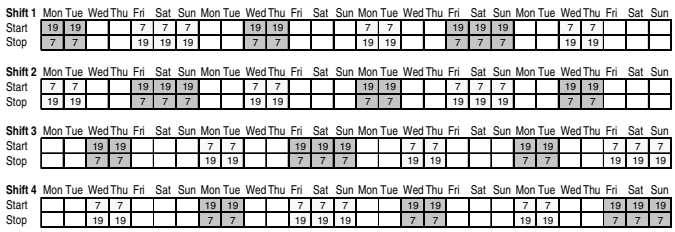


Fig. 3. Illustrative shift schedule

"performance test" is returned to the optimizer and thus forms a continual feedback loop.

The execution time of the optimization model depends on two factors: the simulation time of each iteration and the number of iterations needed to converge to the optimal state. For the staffing optimization problem, the simulation time is mainly affected by the workload volume for each customer/priority group. This is because the simulation needs to run long enough to make the SLA measurement statistically stable. The number of iterations is mainly determined by the number of service agents, the number of service delivery units, and the number of shift schedules. The combination of them defines the optimization space. In our experience the convergence time are normally on the order of hours. Since staffing optimization is typically part of strategic planning that occurs at a much slower time scale (usually on the order of months), the convergence time is of limited concern.

IV. EXPERIMENTAL STUDIES

We implemented and deployed our optimization model at a large service delivery organization. Our deployment covered approximately 300 service functional units including 1,000 service delivery units and 8,000 service agents in 6 service function areas. These service functional units are located in ten geographic regions, with multiple service delivery centers in each region.

Our implementation is built on top of the AnyLogic simulation software, which supports agent-based, system dynamics, and discrete event simulation methodologies in an visual development environment [14], [15]. It also provides several optimization packages (including Scatter Search and Tabu Search that we are using) for Monte Carlo simulations. Using AnyLogic, we specify and code appropriate constraints and performance criteria to yield a reasonable feasible region and to quickly converge to an implementable solution. The developed model is exported as a standalone Java application for users to run models independently.

In this section we provide an experimental evaluation to illustrate how the proposed optimization model can be be used to produce staffing decisions in a services delivery organization. Note that the data has been altered to preserve data privacy and simplified for the illustration purpose, though the nature of day-to-day service operations has been maintained.

A. Delivery Scenario

The studied service functional unit services three types of service requests that form three priorities: console alerts, problem tickets, and ad hoc requests. In Table II we provide basic statistics for these requests. The requests differ in their arrival rate distributions, service time distributions, and service level objectives. The console alerts have the highest weekly volume and most stringent service attainment target, and are serviced with the highest priority. The problem tickets come with a low volume but longer service time, and have the medium priority. The ad hoc requests have the highest average and variability in the service time, but are serviced with the lowest priority.

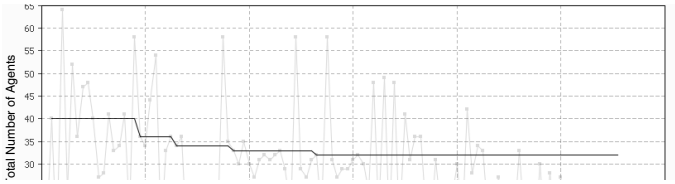
We model the service time using a lognormal distribution. Initially, we model the arrival rate using a stationary Poisson arrival process; afterwards, we use the non-homogeneous Poisson process to capture the varying arrivals for each hour of the week. We use this as an example to demonstrate the need of the simulation model (compared to analytical models) since it can conveniently consider multiple complex factors such as non-homogeneous Poisson process and agent shift schedules.

The service functional unit provides 24 by 7 customer supports due to the requirements from monitoring and handling the console alerts. As shown in Figure 3, the service agents are organized into four shifts and work on a twelve-hour schedule; the the shifts are also "paired" such that, for example, the agents in shift 1 will work from 7am until 7pm followed by agents in shift 2 who work from 7pm to 7am. In addition, management in this service functional unit specifies that a minimum of three agents are required during any hour of the day. This is to accommodate the physical locations of the console monitors which are not collocated at the same place.

B. Optimization Results

The objective of the optimization is to determine the minimal number of agents and ensure that service levels are met for each category of service requests. For simplicity of illustration, we assume this service functional unit only has a single service delivery unit (so that the feasible solutions x_{jk} can be graphed and compared easily).

Figure 4 displays the convergence path of the optimization model. The x-axis indicates the number of iterations (i.e., the number of solutions that have been created and evaluated) and the y-axis indicates the total number of agents in each solution. The grey line in the background shows all candidate



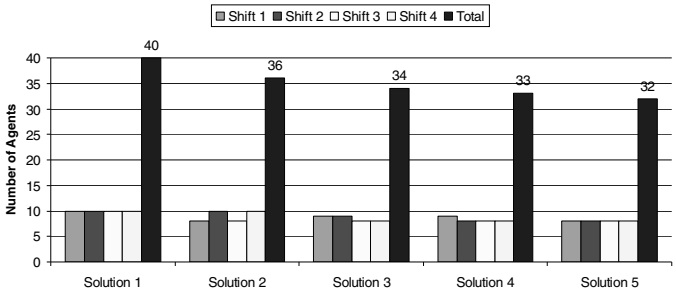


Fig. 5. Number of agents recommended in each feasible solution.

solutions from scatter search and tabu search (including both feasible and infeasible solutions), the line in the middle shows the evolution of the best feasible solutions (i.e., the ones that meet all constraints especially the service level constraints), and the line at the bottom shows the minimum (but infeasible) solutions that have been evaluated. The model starts from a random configuration and finds the first feasible solution at the second iteration (it happens in this example, but generally the first feasible solution may be found much later). The model converges after 102 iterations, during which 53 solutions have been evaluated and five of them are feasible solutions. A solution only needs to be evaluated if the total number of agents from this solution is smaller than that from the current best feasible solution (i.e., below the line in the middle), even though the ones above the line also need to be generated in order to construct the generalized path.

Figure 5 shows the per-shift staffing levels for the five feasible solutions found during this optimization process. Each cluster of bars represents the number of agents recommended for each shift, and the fifth column indicates the total number of agents. The last cluster shows the optimal staffing configuration with 32 agents distributed evenly across four shifts (i.e., eight agents per shift). The balanced shift assignments reflect the stationary arrival of workload that we artificially assumed in the beginning of this section. Further, the agents are, on average, fully utilized over all of their available working hours.

Next, we use the non-homogeneous Poisson process to capture the varying arrivals for each hour of the week. Figure 6 illustrates the volume of arrivals per hour of the week. The x-axis indicates the hours in the week where hour 1 is the hour between Sunday midnight and 1 am Monday morning. The y-axis indicates the total volume of alert requests for

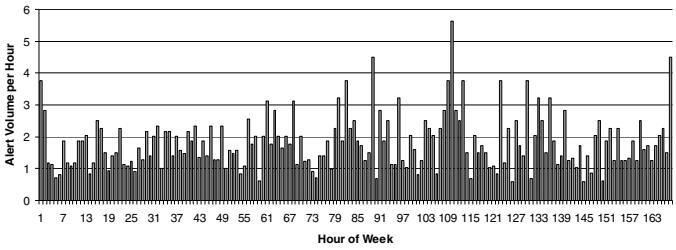


Fig. 6. Volume of alerts per hour of the week

TABLE III
ALTERNATIVE SLA ATTAINMENT SCENARIOS.

	Alerts	Problems	Ad Hoc	Staffing
Base Scenario	99%	95%	95%	32
SLA Scenario 1	99%	99%	99%	32
SLA Scenario 2	95%	95%	95%	29
SLA Scenario 3	90%	90%	90%	29

the corresponding hour of the week. We run the optimization model similarly to the last scenario. However, the new optimal solution requires 64 service agents, twice of that under the stationary arrival scenario. Examination of the staff utilization reveals low average utilization of 52.2% across all shifts. Apparently, the peaks and valleys as observed in Figure 6, even though not significant, do have a drastic impact on the staffing level. This is because of the short target response times and strict service attainment targets, so that many more agents are required to ensure that the service level agreements can be met at the peak of the week.

C. What-if Analysis

Besides staffing level recommendation, we can also use the proposed optimization model to perform what-if analysis and explore alternative service designs. We demonstrate this through an example of how changes in service attainment targets impact the minimum staffing levels. Table III provides four scenarios where the service attainment targets are defined for each service request type. The base scenario is the one studied in Section IV-A, and the other three scenarios provide alternative SLA designs. The last column indicates the minimum number of service agents suggested by the optimization model.

As shown in Table III, when the service attainment level is 99% for all request classes (Scenario 1), the required minimum staffing is the same as that for the base scenario (where the problem and ad hoc requests are taking a less stringent target of 95%). On the other hand, when the attainment level drops to 95% or 90% for all request classes (Scenario 2 and 3), 29 minimum service agents are required in both cases.

The relative staffing level insensitivity in response to attainment level changes is driven by two factors. First, the staffing level is mainly determined by the dominant request type (if it exists). In our example, the dominant request types is the console alert workload which has both the highest volume and

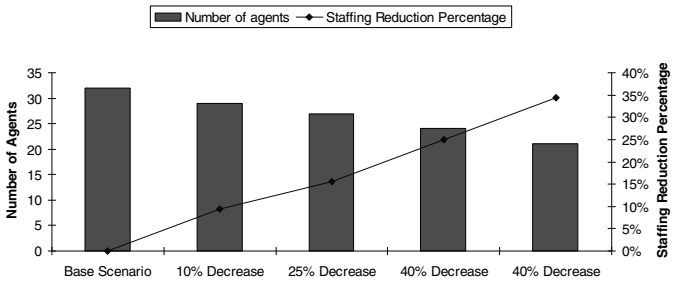


Fig. 7. Required agents for different scenarios with decreased workload volume

the most stringent service level objectives. Thus, increasing the required attainment for problems and ad hoc requests from 95% (in Base Scenario) to 99% (in Scenario 1) has no impact on the staffing levels as compared with the base scenario. Second, due to the discrete nature of the problem, the staffing configuration is restricted to only integer numbers of agents. This explains the noticeable staffing decrease from the Base Scenario to Scenario 2, but no further decrease from Scenario 2 to Scenario 3 (both of them require minimum 29 service agents).

While the above qualitative interpretation can be made without the model, running the optimization model helps to provide quantitative suggestions on service design. For example, in contract negotiation knowing where the attainment level changes make a difference (e.g., between the first two and last two scenarios) or not can help the service designer make informed decisions on the acceptable service attainment target without incurring unnecessary delivery cost.

Next, we consider another example on how automation can benefit the service provider. Use of automation in place of human agents is being explored recently as an effective means to reduce the service delivery cost, especially for the workload that can be handled using standard responses. For the example considered in this section, the console alert workload is the most standard type of service requests where most responses follow standard procedures. Introducing automated tooling for alert handling can decreases the volume of workload that requires response by the service agents. We consider four scenarios where the request volume decreases by 10%, 25%, 40%, and 50% compared to the base scenario defined in Section IV-A. Figure 7 shows the minimum number agents and the staffing reduction percentage in these scenarios, which ranges from 9% to 34%, respectively.

It is worthy to mention that while the above what-if analysis shows significant impact on the required staffing level, on the other hand, many other scenarios we have explored didn't show noticeable difference. Examples include changes in the target response time and changes in the average service time (to mimic the benefit of additional staff training). These observations are not generic across all service functional units, but demonstrate the practical value for using the optimization model for case-dependent analysis.

V. Conclusions and Future Work

The services delivery business is highly dynamic and highly competitive, with thin profit margins. Strict service quality targets coupled with highly variable service request arrival patterns and ever increasing cost containment targets make it challenging for a service delivery provider to deliver consistent quality and remain profitable. Due to various Lean initiatives, there is little room for the provider to pilot alternative solutions that may or may not result in improvements in system performance (reduced cost, improved service quality, etc.) In this paper we proposed an optimization model to provide recommended staffing levels in a complex service delivery system. The optimization model minimizes the total staffing related variable cost while considering the contractual service level constraints, the skills required to respond to different types of service requests, and the shift schedules that the service agents must follow. Given the complexity of service delivery systems, while many formulations of the staffing optimization problem are possible (briefly discussed in III.B), we designed our formulation with the main objectives to be practical and applicable: it covers major concerns in service delivery, requires a reasonable set of parameters, and builds the solution approach that can converge quickly.

The optimization model can be used to determine optimal staffing level. To ensure the model captures the service delivery reality, we conducted model verification and validation using historical data, consulted with the delivery center Subject Matter Experts for modeling assumptions, and piloted solution implementation to verify model recommendations. Furthermore, to improve the usefulness of the model, we conducted real time timing studies in addition to historical workload data, built robustness into the model by considering workload variation, and set in the deployment plan to rerun the model if significant future workload changes occur. Besides optimal staffing, the optimization model can also be used to perform various what-if analysis. For example, a services delivery provider can utilize this optimization framework to measure the impact of adding or removing customers from a service functional unit. Or, it can be used to evaluate the trade-offs of agent training to increase the skill profile versus hiring new agents with a limited set of specific skills.

While the initial results are encouraging, there are various challenges ahead of us. For example, one challenge that we face with our approach is the dynamic nature of the service delivery environment. The supported customer base as well as the number of agents experience frequent changes. Consequently, by the time the optimization results are available (after all data collection and analysis phases which can take months in large deployment), the environment may have changed. Our response to this challenge was to implement methods to shorten the modeling cycle time from initial data collection to completion of final results. In addition, due to the massive scale of model deployment, we would like to create a more user-friendly model. This will help to lower the skills required to run the model and expedite the deployment speed and scope.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Brian Eck, David Northcutt, and George Stark, all employed by IBM, for helpful and constructive discussions that helped us improve the quality of the model. In addition, we are indebted to Anatoly Zherebtsov, employed by XJ Technologies, for his assistance in model development.

REFERENCES

[1] N. Gans, G. Koole, and A. Mandelbaum, "Telephone call centers: Tutorial, review, and research prospects," *Management Science*, vol. 5, pp. 79–141, 2003.

[2] Z. Aksin, M. Armony, and V. Mehrotra, "The modern call center: A multi-disciplinary perspective on operations management research," *Production and Operations Management*, vol. 16, pp. 665–688, 2007.

[3] J. Atlason, M. A. Epelman, and S. G. Henderson, "Optimizing call center staffing using simulation and analytic center cutting-plane methods," *Managment Science*, vol. 54, pp. 295–309, 2008.

[4] M. T. Cezik and P. LaEcuyer, "Staffing multiskill call centers via linear programming and simulation," *Management Science*, vol. 54, pp. 310–323, 2008.

[5] 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.

[6] T. R. Robbins and T. P. Harrison, "A simulation based scheduling moel for call centers with uncertain arrival rates," in *Proceedings of the 2008 Winter Simulation Conference, Miami, FL*, 2008, pp. 2884–2890.

[7] 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.

[8] 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.

[9] 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.

[10] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, "Statistical analysis of a telephone call center: A queueing-science perspective," *Journal of the American Statistical Association*, vol. 100, pp. 36–50, 2005.

[11] R. Ulrich and J. Miller, "Information processing models generating lognormally distributed reaction times," *Journal of Mathematical Psychology*, vol. 37, pp. 513–525, 1993.

[12] F. Glover, M. Laguna, and R. Mart, "Scatter search," in *Advances in Evolutionary Computation: Theory and Applications*, A. Ghosh and e. S. Tsutsui, Eds. New York: Springer-Verlag, 2003, pp. 519–537.

[13] F. Glover, "Tabu search," *Decision Sciences*, vol. 8, pp. 156–166, 1977.

[14] Y. G. Karpov, "Anylogic: A new generation professional simulation tool," in *VI International Congress on Mathematical Modeling, Nizni-Novgorog, Russia*, 2004.

[15] XJ Technologies, "http://www.xjtek.com//," 2011.