RC24954 (W1003-017) March 4, 2010
Mathematics

IBM Research Report

Performance Management of IT Services Delivery

Jianying Hu, Yingdong Lu, Aleksandra Mojsilovic,
Mayank Sharma, Mark S. Squillante
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Performance Management of IT Services Delivery

Jianying Hu, Yingdong Lu, Aleksandra Mojsilović, Mayank Sharma, Mark S. Squillante
Mathematical Sciences Department
IBM Thomas J. Watson Research Center
Yorktown Heights, NY, USA

1. INTRODUCTION

Many applications involve resource allocation problems (RAPs) in which different types of resources are used to provide service to various classes of customers at certain time epochs, else the opportunity to serve the customer is lost. Stochastic loss networks (SLNs) are often used to capture the dynamics and uncertainty of this class of RAPs. Examples include applications in telephony and information technology (IT) networks [13, 18]. Another emerging application is workforce management, and in particular managing the performance of services operations. With the growth of the services sector over the past 50 years, the ability to manage human resources and skills more effectively and efficiently continues to be the critical driver of success for any services company. This is particularly true for IT service providers who typically offer a broad range of service products, each requiring resources with certain capabilities, in markets characterized by highly volatile and uncertain customer demands. Hence, services companies seek innovative solutions to better manage and plan their per-class resource capacity levels in a way that will maximize business performance.

While mathematical models of traditional manufacturing and logistics systems have been developed and used for business performance optimization over the past several decades, these methods cannot be applied directly to related problems in services industries. Human resources are far more complex to model than machines and parts, calling for novel methods that will better capture and represent such complexities. In this paper, we propose a set of mathematical models and an end-to-end solution for performance management in the delivery of IT services. Our goal is to maximize the business performance of an IT services delivery supply chain, starting with the forecasting of the demand for service products and their resource requirements, followed by a form of risk-based capacity planning and the forecasting of multi-skill resource supply, up to the optimal resource assignment to capacity planning targets.

The ultimate objective in all service delivery operations is to have the right people, in the right place, at the right time [1]. A critical first step in achieving this goal is demand forecasting, or more precisely the ability to predict the amount of work that will materialize over the planning horizon, in terms of revenue to the provider, number of engagements to be served, and resource and skill requirements. Service product offerings are usually described only in terms of revenue, duration and solution characteristics, without linkages to resource requirements. In order to obtain a more accurate view of resource demand, we utilize statistical and machine learning methods to determine standardized staffing models for each service product, and then apply these methods and models to estimate the demand and resource capacity requirements for each service product. §2 describes our demand forecasting models.

The uncertainty in engagement demand, process delivery and resource supply is one of the fundamental characteristics of the services business. Therefore, a good understanding of future resource needs and the implications of different resource levels on business performance is essential for success in the marketplace. Service products typically require multiple resources, each capable of employing different skills. Having insufficient resources with the appropriate skills to carry out an engagement when needed on the one hand, or having too many under-utilized resources on the other hand, both result in the loss of profits to the business. To address this revenue-cost tradeoff, we introduce in §3 a risk-based capacity planning approach based on SLNs to determine the capacity targets for each supply resource skill that maximize business performance under the inputs from our demand forecasting models.

In contrast to tactical planning where horizons are short (hours, days, weeks), the timeframes for strategic planning of service delivery to maximize business performance typically span several months, quarters or even a few years. Moreover, service delivery environments are often categorized by significant resource dynamics, such as many employees acquiring skills and changing positions, some employees leaving and new employees being added, all resulting in substantial changes in resource supply characteristics over time. To achieve the forecasting of supply, we develop in §4 stochastic models of the evolution of workforce dynamics over time based on a given set of business policies and economic conditions, which yield estimates for future supply resources and their skill mix.

As the final planning step, §5 presents how desired capacity targets and supply forecasts are taken as inputs to our gap/glut optimization to determine the optimal assignment of available multi-skill supply resources and compute the gaps and gluts across all skills under this assignment. A variation of our risk-based capacity planning is used to determine the optimal hiring, training and retention actions to address per-skill gaps and gluts.

Our end-to-end planning solution has been applied extensively to support several service delivery units within the IBM Corporation. We present in §6 a representative sample of numerical experiments from the use of our system with real data from the IT services industry. Some of our experiences with this general performance management approach are also discussed. Additional technical details and mathematical results as part of our end-to-end solution can be found in the references provided throughout the paper.

2. DEMAND FORECASTING MODELS

Demand forecasting is the important first step in the performance management of IT services delivery. Service product engagements typically require multiple resources, each having different skills. The total demand for these skills over any future time period is a combination of demand coming from three sources: (1) *ongoing engagements*, i.e., engagements that have already started and are

being delivered; (2) *opportunities*, i.e., potential deals in the sales pipeline at different sales stages; and (3) *expected deals*, i.e., deals of various types that are expected based on market research and experience, but are not yet concrete enough to be entered into the sales pipeline. A key component for the accurate forecasting of all three sources of demand is the standard staffing models that link engagement types and expected revenue to skill requirements.

2.1 Automated Generation of Staffing Models

In order to obtain a more accurate view of resource and skill requirements and to better forecast resource demands, given the missing linkage between service product offerings/opportunities and required resources, we need standardized staffing models that can be used to determine the expected staffing requirements based on each solution type. Such models allow for effective planning of staffing decisions at earlier stages of the engagement process, more reliable forecasting of resource needs, and better workforce planning. This model-based approach calls for the creation of an engagement, or project, categorization scheme, which would link a set of project attributes to typical resource requirements over the project life cycle. Such an approach requires the development of a systematic method for creating a solution taxonomy, and estimating staffing models (i.e., the specification of staffing needs in terms of required hours of each skill each week for the planned project) automatically, on the basis of key engagement characteristics.

We have developed a methodology based on statistical clustering techniques for generating groups of similarly staffed projects, using information on reported labor hours from a large number of historical projects from the enterprise labor-claim management system [7]. The approach utilizes a variant of the hierarchical-k-means algorithm proposed in [6] to identify homogeneous groups of projects with respect to the resource utilization vectors. Once the statistical cluster analysis is complete, we then create a representative solution taxonomy by: (1) examining the distribution of values of project attributes in each cluster; (2) creating an appropriate name and description for each cluster; and (3) validating each cluster assignment and refining taxonomy labels and class descriptions through discussions with subject matter experts.

Once a standard solution taxonomy has been created, it will likely need to be re-examined and adjusted on a regular basis due to highly dynamic business environments and shifting customer needs. This iterative process requires accurate tracking of projects according to the standard taxonomy to provide critical business insights. Examples of such insights directly relevant to resource demand forecasting include: whether each identified solutions type is being delivered as planned; if not, whether the staffing model needs to be adjusted accordingly; whether a steady shift in delivery patterns can be detected which may indicate the need for a new solution; etc. Because the solution taxonomy is constantly evolving through this iterative process, and due to the inherent need of some degree of customization during delivery, it is often very difficult for project managers to categorize projects accurately and comprehensively. Hence, there is a need for an automated methodology to assist with project categorization during actual delivery. We have developed such a methodology by formulating the problem in a semi-supervised clustering framework [8, 9]. In the first step of this approach, project descriptions are matched against the solution category descriptions. For a subset of the data, this step produces a category label, along with a confidence score for each sample point. In the second step, the set of labeled data along with the confidence scores are used as soft seeds in a semi-supervised clustering algorithm to categorize all projects using a semi-supervised version of the k-mean algorithm called "soft seeded k-means". We refer the reader to [8] for more detailed descriptions of this algorithm.

The process of automated generation and adjustment of staffing models as described above is carried out on a regular basis to allow for an evolving solution portfolio that reflects changes in market and customer requirements. The resulting dynamic project taxonomy along associated staffing models provide critical input to ensure accurate forecasting for resource demands coming from all three major sources. For ongoing engagements, they allow for insight into future resource requirements and expected roll-off dates for already assigned resources. For opportunities and expected deals, they provide the necessary linkage between revenue from different types of solutions and resource requirements.

2.2 Integrated Demand Forecasting System

An integrated demand forecasting system was developed to account for demands from all three major sources. For ongoing engagements, up-to-date delivery information is used to compute the expected number of ongoing projects over future periods, as well as necessary customizations to standard staffing templates, and expected roll-off dates of already deployed resources. For opportunities in the pipeline, statistical regression models are used to predict the win probability of each deal in the pipeline based on attributes such as lapse time and recent movement in pipeline, customer information, deal size, etc. These win probabilities are then used to compute the expected number of engagements for each solution type (with its associated staffing templates) in future periods. For expected deals, quarterly revenue targets are used to provide an engagement level estimate. More specifically, for each solution type, the difference between the revenue target and the expected revenue from ongoing engagements and pipeline opportunities is considered to be the total expected revenue from expected deals. The typical deal size for this type of engagement is then used to compute the corresponding expected number of engagements. Whenever there is inconsistency in the information, e.g., when the expected revenue coming form the opportunities and ongoing engagements exceeds the revenue target, or the expected deal portion is exceptionally large, a management exception is raised, and all stakeholders (i.e., leaders from sales, delivery and planning) are engaged to make necessary adjustments to resolve the inconsistencies.

3. RISK-BASED CAPACITY PLANNING

The estimates of per-product demand and per-skill capacity requirements for each service product from our forecasting models are provided as input to the capacity planning component of our end-to-end process. For ongoing engagements already being delivered, the existing assignments of per-skill resource capacities are known and can be directly included in the capacity plan. However, when the duration of product engagements are relatively shorter than the overall time horizon of interest, then the sales pipeline opportunities and expected deals represent an important fraction of the per-product demand. Moreover, resource capacities assigned to ongoing engagements will become available as these engagements are completed. To capture these various sources of stochastic per-product demand and per-skill capacities, we model the IT services delivery capacity planning problem as a SLN in which the risk of losing service product engagements due to insufficient supply resources in one or more required skills at the time the service product needs to be delivered is captured in the stationary loss probabilities of the SLN. Here, the multi-class stochastic arrival process is used to represent the time epochs at which the different product engagements must be delivered. The overall time horizon consists of a sequence of coarse subintervals, each involving a stationary SLN under a fixed set of parameters that changes from one subinterval

to the next according to a general Markov-modulated process.

Then the objective of our capacity planning optimization problem is to determine the per-skill resource capacity levels for this SLN that maximize the expectation of a business performance utility function over a long-run time horizon. We shall assume that the length of each subinterval is sufficiently long for the multidimensional stochastic process modeling the loss network to reach stationarity, where the multiple time scales involved in IT services delivery provide both theoretical and practical support for our stationary stochastic approach. The utility function is based on rewards gained for delivering IT service products that can be serviced at the time of their required delivery instant (arrival) and on penalties incurred as the result of deploying resource capacity levels over the multi-period time horizon. Constraints on the stationary loss-risk probability for each service product can be added to the optimization formulation in order to guarantee a specific level for the corresponding acceptance rate, or serviceability level.

In the interest of space, single-period versions of our stochastic models and stochastic optimization will be presented below. We refer the interested reader to [5] for a description of a multi-period version of our stochastic models and stochastic optimization, together with related technical details and mathematical results.

3.1 Stochastic Models

Consider a SLN consisting of a set of skills $\mathcal{L}$ and a set of service products $\mathcal{R}$, with product delivery engagements comprised of collections of per-skill capacities. The delivery of product $r \in \mathcal{R}$ requires $A_{jr} \geq 0$ units of capacity from skill $j \in \mathcal{L}$, where each skill j has $C_j \geq 0$ units of capacity overall. Engagement instances for product r need to be delivered according to an independent Poisson process with rate ν_r. Such a product delivery opportunity is lost if the available capacity for any skill j is less than A_{jr}, and otherwise the product engagement is delivered and reserves capacity A_{jr} for each skill j throughout the duration of the service product delivery. The delivery engagement duration times are i.i.d. following a general distribution with unit mean (without loss of generality). Product delivery instants and duration times are mutually independent. Let L_r denote the stationary loss probability for product r, and E_j the stationary blocking probability for skill j. Define $\mathbf{A} \triangleq [A_{jr}]_{j=1,\ldots,|\mathcal{L}|;r=1,\ldots,|\mathcal{R}|}$, $\mathbf{C} \triangleq (C_1, C_2, \ldots, C_{|\mathcal{L}|})$, $\boldsymbol{\nu} \triangleq (\nu_1, \ldots, \nu_{|\mathcal{R}|})$, $\mathbf{L} \triangleq (L_1, \ldots, L_{|\mathcal{R}|})$, $\mathbf{E} \triangleq (E_1, \ldots, E_{|\mathcal{L}|})$. Note that the rth column of matrix $\mathbf{A}$ corresponds to the staffing template for service product r from our demand forecasting models.

Let $\boldsymbol{n}(t) = (n_1(t), \ldots, n_{|\mathcal{R}|}(t)) \in \mathbb{Z}_+^{|\mathcal{R}|}$ be the vector of the number of active service delivery engagements in the network at time t. By definition, $\boldsymbol{n}(t) \in \mathcal{S}(\mathbf{C}) = \{\boldsymbol{n} \in \mathbb{Z}_+^{|\mathcal{R}|} : \mathbf{A}\boldsymbol{n} \leq \mathbf{C}\}$. As Erlang originally established, together with extensions by various researchers, it is well known that there is a unique stationary distribution $\boldsymbol{\pi}$ on the state space $\mathcal{S}(\mathbf{C})$ such that

$$\boldsymbol{\pi}(\boldsymbol{n}) \;=\; G(\mathbf{C})^{-1} \prod_{r=1}^{|\mathcal{R}|} \frac{\nu_r^{n_r}}{n_r!} \tag{1}$$

for $\boldsymbol{n} \in \mathcal{S}(\mathbf{C})$, where $G(\mathbf{C})$ is the normalizing constant

$$G(\mathbf{C}) \;=\; \sum_{\boldsymbol{n} \in \mathcal{S}(\mathbf{C})} \prod_{r=1}^{|\mathcal{R}|} \frac{\nu_r^{n_r}}{n_r!}. \tag{2}$$

The stationary probability that a product r engagement is lost can be expressed as $L_r = 1 - G(\mathbf{C})^{-1} G(\mathbf{C} - \mathbf{A}\mathbf{e}_r)$, where $\mathbf{e}_r$ is the unit vector for a single active product r delivery engagement.

Due to the computational complexity of the normalizing constant, known to be $\sharp P$-complete in the size of the network [14], the Erlang fixed-point approximation (EFPA) has been long used as a more efficient alternative to the exact Erlang loss formula. The EFPA is based on approximating the stationary blocking probabilities of the individual skills, E_j, by a set of fixed-point equations:

$$\rho_j \;=\; (1 - E_j)^{-1} \sum_{r=1}^{|\mathcal{R}|} A_{jr} \nu_r \prod_{i=1}^{|\mathcal{L}|} (1 - E_i)^{A_{ir}}, \tag{3}$$

$$E_j \;=\; \mathcal{E}\left(\rho_j, C_j\right), \qquad \mathcal{E}(\nu, C) \;=\; \frac{\nu^C}{C!} \left(\sum_{n=0}^{C} \frac{\nu^n}{n!} \right)^{-1}, \tag{4}$$

where the last expression is the Erlang formula for the loss probability of an isolated skill with capacity C under arrival rate ν. Then the stationary loss probability for service product r can be approximated in terms of the per-skill blocking probabilities as

$$L_r \;=\; 1 - \prod_{j=1}^{|\mathcal{L}|} (1 - E_j)^{A_{jr}}. \tag{5}$$

This approximation assumes that lost product delivery engagements are caused by independent blocking events on each of the skills comprising the products, using the one-dimensional Erlang loss formula for each skill with an appropriately thinned arrival rate. We refer the interested reader to [13] for additional details.

It is well known that there exists a unique solution $\mathbf{E} \in [0, 1]^{|\mathcal{L}|}$ of the EFPA equations (4). Kelly [13] further established that this EFPA solution converges to the exact solution of the Erlang loss model (ELM) in a large network limiting regime (LNLR) where the arrival rates and resource capacities are increased in a proportional manner with respect to a scaling parameter $N \in \mathbb{N}$: $\mathbf{C}_N = N\mathbf{C} = (NC_1, \ldots, NC_{|\mathcal{R}|})$, $\boldsymbol{\nu}_N = N\boldsymbol{\nu} = (N\nu_1, \ldots, N\nu_{|\mathcal{R}|})$. The asymptotic exactness of the EFPA follows from an instance of the central limit theorem for conditional Poisson r.v.s in which $\boldsymbol{\nu}$ and $\mathbf{C}$ grow together. Namely, the $|\mathcal{R}|$ Poisson r.v.s being truncated by a polytope involving the capacities $\mathbf{C}$ are approximated by $|\mathcal{R}|$ independent normal r.v.s truncated by the polytope.

On the other hand, it is also well known that the EFPA can provide relatively poor estimates for the per-product loss probabilities L_r in various model instances. To address this, one can consider the ELM as a stable system where admitted product delivery engagements experience an average delay of 1 and lost delivery engagements experience a delay of 0. Hence, the average delay experienced by delivery engagements for product r is given by

$$D_r = (1 - L_r) \times 1 + L_r \times 0 = (1 - L_r),$$

which together with Little's law yields

$$1 - L_r \;=\; \frac{\mathbb{E}[n_r]}{\nu_r}, \tag{6}$$

and thus L_r can be obtained through $\mathbb{E}[n_r]$. By definition, $\mathbb{E}[n_r] = \sum_{k=0}^{\infty} k \Pr[n_r = k]$, so $\mathbb{E}[n_r]$ can be obtained through approximations of $\Pr[n_r = k]$, which corresponds to the probability mass along the "slice" of the polytope defined by $n_r = k$. The family of slice methods (SMs) introduced in [11] is based on approximations for $\Pr[n_r = k]$ that assume the mass along each slice is concentrated around the mode of the distribution restricted to the slice.

From the definition of the stationary distribution $\boldsymbol{\pi}(\cdot)$, the mode $\boldsymbol{n}^*$ corresponds to a solution of the optimization problem

$$\max \;\; \sum_r n_r \log \nu_r - \log n_r! \qquad \text{over} \qquad \boldsymbol{n} \in \mathcal{S}(\mathbf{C}).$$

A natural continuous relaxation of the state space $\boldsymbol{n} \in \mathcal{S}(\mathbf{C})$ is

$$\bar{\mathcal{S}}(\mathbf{C}) = \{\mathbf{x} \in \mathbb{R}_+^{|\mathcal{R}|} : \mathbf{A}\mathbf{x} \leq \mathbf{C}\},$$

for which we obtain the corresponding optimization problem (P1):

$$\max \sum_r x_r \log \nu_r - \log \Gamma(x_r + 1) \qquad \text{over} \qquad \mathbf{x} \in \bar{\mathcal{S}}(\mathbf{C}).$$

For each value of $k \in \{n_r : \mathbf{n} \in \mathcal{S}(\mathbf{C})\}$ that is along each *slice*, define $\mathbf{x}^*(k, r)$ to be the solution of the optimization problem (P2):

$$\max \sum_r x_r \log \nu_r - \log \Gamma(x_r + 1) \quad \text{over} \quad \mathbf{x} \in \bar{\mathcal{S}}_{k,r}(\mathbf{C})$$

where $\bar{\mathcal{S}}_{k,r}(\mathbf{C}) \equiv \bar{\mathcal{S}}(\mathbf{C}) \cap \{\mathbf{x} : x_r = k\}$. A computationally implementable version of (P1) and (P2) can be obtained by Stirling's approximation and ignoring the $O(\log n_r)$ term, which respectively yields the following convex relaxations for (P1), (P2):

$$\max \sum_r x_r \log \nu_r + x_r - x_r \log x_r \quad \text{over} \quad \mathbf{x} \in \bar{\mathcal{S}}(\mathbf{C}), \quad (7)$$

$$\max \sum_r x_r \log \nu_r + x_r - x_r \log x_r \quad \text{over} \quad \mathbf{x} \in \bar{\mathcal{S}}_{k,r}(\mathbf{C}). \quad (8)$$

In the general SM [11, 12], for each product r, problem (7) is solved for the mode of the distribution $\mathbf{x}^*$ and problem (8) is solved for each slice defined by $n_r = k$, $k \in \{n_r : \mathbf{n} \in \mathcal{S}(\mathbf{C})\}$. To reduce this computational complexity and provide computational complexity similar to that of the EFPA, a 3-point SM is also presented in [11, 12]. Instead of computing $\mathbf{x}^*(k, r)$ for all $k \in \{n_r : \mathbf{n} \in \mathcal{S}(\mathbf{C})\}$, the 3-point SM consists of solving (8) for $k = 0$ and the maximum value of k, and also obtaining the mode $\mathbf{x}^*$ by solving (7). Then $\mathbf{x}^*(k, r)$ is approximated for all other values of k by linear interpolation between pairs of the 3 computed modes. Denoting the objective function from (7) and (8) by

$$q(\mathbf{x}) = \sum_r x_r \log \nu_r + x_r - x_r \log x_r,$$

the estimate of $\mathbb{E}[n_r]$ is obtained as

$$\mathbb{E}[n_r] \;=\; \frac{\sum_k k \exp(q(\mathbf{x}^*(k, r)))}{\sum_k \exp(q(\mathbf{x}^*(k, r)))}, \quad (9)$$

from which we calculate (6). It is established in [11, 12] that the SMs are asymptotically exact in the same LNLR considered by Kelly and provide superior accuracy over the EFPA, in general as well as especially in the critically loaded regime [10].

For more details on the SMs for approximating the stationary loss probabilities in the ELM, we refer to [11, 12]. A refinement of the SMs and a new randomized contour method to approximate both the stationary distribution and the loss probabilities are presented in [2]. Methods for approximating the stationary loss probabilities in SLNs under general renewal arrival processes are developed in [15, 16], together with a continuous relaxation of the SLN to accommodate resource capacities as continuous variables.

3.2 Stochastic Optimization

Consider a stochastic optimization problem to determine the capacity vector $\mathbf{C}^*$ of a SLN with capacity requirement matrix $\mathbf{A}$ and arrival rate vector $\boldsymbol{\nu}$ that maximizes a profit function over a long-run time horizon, where revenues are gained for accepted product deliveries and costs are incurred for deployed capacities. Formally, the objective function for a main formulation of interest is given by

$$\max_{\mathbf{C}} \quad \sum_{r=1}^{|\mathcal{R}|} u_r \left(1 - L_r\right) \nu_r \;\; - \;\; \sum_{j=1}^{|\mathcal{L}|} v_j \, C_j, \quad (10)$$

where u_r is the base revenue rate for product r deliveries and v_j is the base cost rate for skill j capacity. The constraints of the problem include equations (6) & (9) or (3) – (5) depending upon whether the SM or the EFPA are being used to obtain the stationary loss probabilities L_r of the SLN. Furthermore, constraints on the per-product loss probabilities of the form $L_r \leq \beta_r$, for $\beta_r \in [0, 1)$, can be added to the formulation to guarantee certain levels for the fractions of accepted product deliveries which, in turn, are also related to the per-product serviceability levels and market share. Alternative formulations of interest include maximizing revenue or minimizing cost subject to constraints on the gross profit margin.

It is known that the solution of the optimization problem (10) based on the EFPA (3) – (5) converges asymptotically to the solution of the optimization problem (10) based on the exact ELM (1) – (2) in the LNLR; see, e.g., [19, 13]. We can also establish the corresponding result for our SM approximations, namely that the solution of the optimization problem (10) based on our SMs (6) & (9) converges asymptotically to the solution of the optimization problem (10) based on the exact ELM (1) – (2) in the LNLR.

The objective (10) is based on the expected total profit rate where the first summation represents the expected total revenue rate and the second summation represents the total cost rate, both as a function of the capacity vector decision variable $\mathbf{C}$ given a capacity requirement matrix $\mathbf{A}$, an arrival rate vector $\boldsymbol{\nu}$, and base revenue and cost rates $\mathbf{u} \triangleq (u_1, \ldots, u_{|\mathcal{R}|})$, $\mathbf{v} \triangleq (v_1, \ldots, v_{|\mathcal{L}|})$. Although the total cost rate is linear in $\mathbf{C}$, the expected total revenue rate is a nonlinear function of the capacity vector since $\mathbf{L}$ is a nonlinear function of $\mathbf{C}$. Large-scale nonlinear programming (NLP) problems can be extremely complex and difficult to solve in an efficient manner; see, e.g., [3]. We therefore exploit the properties of the NLP (10) based on our results for (6) & (9) or (3) – (5) together with state-of-the-art large-scale NLP solvers to compute the optimal capacity vector $\mathbf{C}$ in a very efficient manner for values of $|\mathcal{L}|$ and $|\mathcal{R}|$ on the order of several hundreds or thousands.

4. SUPPLY FORECASTING MODELS

Another component of our end-to-end performance management process consists of stochastic models and methods for forecasting available supply resources and their capabilities (skills) over the time horizon of interest. This involves estimating the evolution of workforce dynamics over time, including future hiring, future attrition and future transitions within the existing workforce, based on historical information and a given set of business policies and economic conditions. As economic conditions are uncertain and can quickly change, our stochastic models and methods support interactive sessions and scenario analysis to develop robust forecasts of future supply resources and skill mix. Since business policies and actions can be changed and future workforce dynamics can be influenced in desired directions, we have also developed solutions for the corresponding stochastic optimization problem to determine how the workforce should evolve over time to maximize business performance. These stochastic optimization and related results are omitted due to space restrictions, and we refer the interested reader to [17] for a description of such stochastic evolution optimization capabilities as well as for more details on the stochastic models and methods to estimate the evolution of workforce dynamics over time. The results of our supply forecasting models and methods are then provided as input to our multi-skill gap/glut optimization.

4.1 Stochastic Evolution Models

Recall that $\mathcal{L}$ denotes the set of skills labelled by j. Since each resource comprising the supply is capable of employing a subset of different skills, let $\mathcal{J}$ denote the family of subsets of the set of skills $\mathcal{L}$ that are possessed by supply resources, labelled by i. Define $\mathbf{y}(t) \equiv (y_1(t), \ldots, y_{|\mathcal{J}|}(t))$ to be the workforce state vector

where $y_i(t)$ denotes the expected number of resources of type i at time t, $i \in \mathcal{J}$, $t = 0, \ldots, T$, and T is the time horizon of interest. The hiring vector $\mathbf{h}(t) \equiv (h_1(t), \ldots, h_{|\mathcal{J}|}(t))$ denotes the expected number of hires for every resource type and the attrition vector $\mathbf{a}(t) \equiv (a_1(t), \ldots, a_{|\mathcal{J}|}(t))$ denotes the expected amount of attrition for every resource type, both defined for each time interval $[t, t+1)$ in order to model the workforce dynamics at the desired temporal granularity. Examples of time intervals of interest can include weekly, monthly, quarterly and yearly. Further we use component-wise division to define the corresponding hiring rate and attrition rate vectors $\boldsymbol{\lambda}(t) \equiv \boldsymbol{h}(t)/\boldsymbol{y}(t)$ and $\boldsymbol{\mu}(t) \equiv \boldsymbol{a}(t)/\boldsymbol{y}(t)$.

Let $p_{ii'}(t)$ denote the probability that a type-i labor resource transitions to become a type-i' labor resource over the time interval $[t, t+1)$, where $\sum_{i' \in \mathcal{J}} p_{ii'}(t) \leq 1$. When $a_i(t) > 0$ then the inequality is strict (i.e., $\sum_{i' \in \mathcal{J}} p_{ii'}(t) < 1$) and $1 - \sum_{i' \in \mathcal{J}} p_{ii'}(t)$ represents the probability that a type-i labor resource leaves the workforce. The corresponding workforce evolution one-step transition probability matrix is given by $\mathbf{P}(t) \equiv [p_{ii'}(t)]_{i,i' \in \mathcal{J}}$. Finally, let the workforce cost vector $\mathbf{c}(t) \equiv (c_1(t), \ldots, c_{|\mathcal{J}|}(t))$ denote the expected cost (e.g., salaries, benefits) of the labor resources at time t. Note that by making the transition probability matrices and model vectors a function of each time interval t, our stochastic models support time-varying behaviors of various forms (including seasonal effects) for the evolution of workforce dynamics.

There are a wide variety of approaches available to set the parameters of our stochastic workforce evolution models. One relatively simple and straightforward approach can be based on extracting the base model parameters, such as transition probabilities, from available historical workforce data in the following manner. The elements of $\boldsymbol{P}(t)$ are simply calculated as

$$p_{ii'}(t) \;=\; \frac{|i \to i'|(t)}{\sum_{j=1}^{|\mathcal{J}|} |i \to j|(t) + a_i(t)},$$

which ensures that

$$\left(1 - \sum_{i'=1}^{|\mathcal{J}|} p_{ii'}(t)\right) \;=\; \frac{a_i(t)}{\sum_{j=1}^{|\mathcal{J}|} |i \to j|(t) + a_i(t)},$$

where $|i \to i'|(t)$ denotes the number of transitions from state i to state i' over the time interval $[t, t+1)$, for all $i, i' \in \mathcal{J}$. Other model parameters can be calculated in an analogous manner.

Over each time interval, the net dynamics for labor resources of type-i are comprised of

- $y_i(t)$: the number of labor resources in state i at the beginning of the time interval;
- $h_i(t)$: the flow of labor resources into state i due to hiring over the time interval;
- $\sum_{i' \neq i} y_{i'}(t) p_{i'i}(t)$: the transitions into state i from other labor resource types over the time interval;
- $y_i(t) \sum_{i' \neq i} p_{ii'}(t)$: the transitions to other labor resource types from state i over the time interval; and
- $y_i(t)(1 - \sum_{i'} p_{ii'}(t))$: the outflow of labor resources from state i due to attrition over the time interval.

More formally, we have

$$y_i(t+1) \;=\; y_i(t) + h_i(t) + \sum_{i'=1:i' \neq i}^{|\mathcal{J}|} y_{i'}(t) p_{i'i}(t) - y_i(t) \sum_{i'=1:i' \neq i}^{|\mathcal{J}|} p_{ii'}(t) - y_i(t)(1 - \sum_{i'=1}^{|\mathcal{J}|} p_{ii'}(t)),$$

which simplifies to $y_i(t+1) = h_i(t) + \sum_{i'=1}^{|\mathcal{J}|} y_{i'}(t) p_{i'i}(t)$, or in matrix form

$$\mathbf{y}(t+1) \;=\; \mathbf{h}(t) + \mathbf{y}(t)\mathbf{P}(t). \tag{11}$$

By iterating (11) it follows that, for every $s = 1, \ldots, T$,

$$\mathbf{y}(s) \;=\; \sum_{t=0}^{s-1} \mathbf{h}(t) \prod_{t'=t+1}^{s-1} \mathbf{P}(t') + \mathbf{y}(0) \prod_{t=0}^{s-1} \mathbf{P}(t),$$

from which we obtain the terminal workforce state vector

$$\mathbf{y}(T) \;=\; \sum_{t=0}^{T-1} \mathbf{h}(t) \prod_{t'=t+1}^{T-1} \mathbf{P}(t') + \mathbf{y}(0) \prod_{t=0}^{T-1} \mathbf{P}(t).$$

Equivalently, representing the dynamics in terms of rates, we have

$$\mathbf{y}(t+1) \;=\; [\boldsymbol{\lambda}(t) * \mathbf{y}(t)] + \mathbf{y}(t)\mathbf{P}(t) \;=\; \mathbf{y}(t)[\boldsymbol{\Lambda}(t) + \mathbf{P}(t)],$$

where $[\boldsymbol{\lambda}(t) * \mathbf{y}(t)] \equiv [\lambda_1(t) y_1(t), \ldots, \lambda_{|\mathcal{J}|}(t) y_{|\mathcal{J}|}(t)] = \mathbf{h}(t)$, $\boldsymbol{\Lambda}(t)$ is a diagonal matrix with $\Lambda_{ii}(t) = \lambda_i(t)$. Iterating yields

$$\mathbf{y}(T) \;=\; \mathbf{y}(0) \prod_{t=0}^{T-1} [\boldsymbol{\Lambda}(t) + \mathbf{P}(t)].$$

Suppose, in addition to the trajectory of the workforce, we also seek to determine how labor costs will evolve. Let $K(t)$ denote the expected total cumulative costs of all labor resources over the time interval $[0, t)$, $t = 1, 2, \ldots, T$. We then have $K(T) = \sum_{t=1}^{T} [\mathbf{c}(t) \cdot \mathbf{y}(t)]$, where $[\mathbf{c}(t) \cdot \mathbf{y}(t)] = c_1(t) y_1(t) + \ldots + c_{|\mathcal{J}|}(t) y_{|\mathcal{J}|}(t)$.

4.2 Scenario Analysis

In order to study the evolution of workforce dynamics under different business policies, economic conditions and other exogenous choices influencing one or more of the model parameters, it can be valuable to analyze the results of our stochastic workforce models under different scenarios. We next consider various forms of scenario (or what-if) analysis based on proportional increases or decreases in certain model parameters resulting from changes in policies, conditions and other choices. More precisely, let $\delta(y)$ denote the proportional change to be made to its argument y. Thus, if we set y to be some model parameter then $\delta(y) < 0$ ($\delta(y) > 0$) would signify a relative decrease (increase) in y. The absolute change in y is then given by $(1 + \delta(y))y$, which henceforth shall be written simply as $(1+\delta)y$. To illustrate the scenario analysis capability, we will focus only on global changes where some parameter is modified by the same proportional change δ for all $i = 1, \ldots, |\mathcal{J}|$ and $t = 0, \ldots, T$. The model parameters can then be updated to accommodate a particular scenario as follows.

Changes in Additions: Let δ_h denote the desired relative change to the hiring numbers $\mathbf{h}(t)$. Then the expression for the new expected number of hires of type-i labor resources over the time interval $[t, t+1)$ is $h_i^{\text{new}}(t) = (1 + \delta_h) h_i(t)$, for all $i = 1, \ldots, |\mathcal{J}|$ and $t = 0, \ldots, T-1$. The corresponding new expected rate of hires of type-i labor resources over $[t, t+1)$ is given by $\lambda_i^{\text{new}}(t) = h_i^{\text{new}}(t)/x_i(t)$.

Changes in Attrition: Let δ_a denote the desired relative change to $\boldsymbol{a}(t)$. Then the new expected amount of attrition of type-i labor resources over the time interval $[t, t+1)$ is given by $a_i^{\text{new}}(t) = (1 + \delta_a) a_i(t)$, for all $i = 1, \ldots, |\mathcal{J}|$ and $t = 0, \ldots, T-1$. To realize the desired overall attrition probability of $(1 - \sum_{i'} p_{ii'}(t))(1 + \delta_a)$,

we adjust the transition probabilities as follows:

$$p^{\text{new}}_{ii'}(t) \;=\; \frac{p_{ii'}(t)}{\sum\limits_{i'} p_{ii'}(t)} \left(1 - (1 - \sum_{i'} p_{ii'}(t))(1 + \delta_a)\right).$$

Changes in Costs: If δ_c denotes the desired relative change to labor resource costs, then the new expected total cumulative costs of labor resources over the time interval $[0, T)$ is given by $K^{\text{new}}(T) = (1 + \delta_c) \sum_{t=1}^{T} [\mathbf{c}(t) \cdot \mathbf{y}^{\text{new}}(t)]$.

5. GAP/GLUT OPTIMIZATION

We next take as input the per-skill capacity targets from our capacity planning component together with the multi-skill supply resources from our supply forecasting models and consider the optimal assignment of the latter to the former. The existing assignments of per-skill resource capacities for ongoing engagements can be subtracted from the per-skill capacity targets and the multi-skill supply resources with the gap/glut optimization focusing on optimally matching the remaining supply and demand. Alternatively, when reassignments are possible from a business perspective, all of the per-skill capacity planning targets and multi-skill supply forecasts are used as input to the gap/glut optimization component of our end-to-end process, thus providing minimal gaps and gluts.

Recall that $\mathcal{L}$ denotes the set of skills labelled by j and that $\mathcal{J}$ denotes the family of subsets of the set of skills $\mathcal{L}$ that are possessed by supply resources, which we label by k. Suppose that the capacity target for skill $j \in \mathcal{L}$ is d_j, and suppose that the number of available supply resources with skill subset $k \in \mathcal{J}$ is r_k. We shall use the simplifying notation $j \in k$ to state that $j \in \mathcal{L}$ is an element of the subset $k \in \mathcal{J}$. Let b_j and c_j be the gap and glut for each skill j, respectively, and let α_{jk} be the amount of supply resources capable of employing skill subset $k \in \mathcal{J}$ that are assigned to employ skill $j \in \mathcal{L}$. Then we can use the following linear programming (LP) formulation as a base optimization model for assigning the multi-skill supply resources to the per-skill capacity targets:

$$\begin{aligned}
\min \quad & \sum_{j=1}^{|\mathcal{L}|} (w^b_j b_j + w^c_j c_j) && && (12)\\
s.t. \quad & \sum_{k=1}^{|\mathcal{J}|} \alpha_{jk} + b_j = d_j + c_j, && \forall j \in \mathcal{L},\\
& \sum_{j=1}^{|\mathcal{L}|} \alpha_{jk} = r_k, && \forall k \in \mathcal{J},\\
& \alpha_{jk} = 0, && \forall j \notin k,\ \ \forall k \in \mathcal{J},\\
& \mathbf{b} \geq 0, \mathbf{c} \geq 0, \alpha_{jk} \geq 0,
\end{aligned}$$

where w^x_j and w^c_j are weights for the gaps and gluts associated with skill $j \in \mathcal{L}$, respectively. Note that the solution of this LP also provides the calculation of the gaps and gluts under the optimal assignment of the supply resources to the capacity targets.

In some instances of this problem, we may also want to include preferences among the subsets of skills when assigning multi-skill supply resources to the per-skill capacity targets. Let us introduce an additional family of variables z_{jk} that represent the remaining amount of supply resources capable of employing skill subset $k \in \mathcal{J}$ that are not assigned to employ skill $j \in \mathcal{L}$. This causes the second set of constraints to become $\sum_{j=1}^{|\mathcal{L}|} \alpha_{jk} + z_{jk} = r_k, \forall k \in \mathcal{J}$. Then, whenever subset k is preferred over subset k' for skill j, we add $M z_{jk} + z_{jk'}$ in the objective where M is a large real number as in the big-M method for solving LPs [4].

The resulting extension of our base LP model for multi-skill assignment with preferences can be expressed as:

$$\begin{aligned}
\min \quad & \sum_{j=1}^{|\mathcal{L}|} (w^b_j b_j + w^c_j c_j) + \sum_{j,k,k'} (M_{j,k,k'} z_{jk} + z_{jk'}) && && (13)\\
s.t. \quad & \sum_{k=1}^{|\mathcal{J}|} \alpha_{jk} + b_j = d_j + c_j, && \forall j \in \mathcal{L},\\
& \sum_{j=1}^{|\mathcal{L}|} \alpha_{jk} + z_{jk} = r_k, && \forall k \in \mathcal{J},\\
& \alpha_{jk} = 0, \ \ z_{jk} = 0, && \forall j \notin k,\ \ \forall k \in \mathcal{J},\\
& \mathbf{b} \geq 0, \mathbf{c} \geq 0, \alpha_{jk} \geq 0, z_{jk} \geq 0.
\end{aligned}$$

The solution also provides the gap and glut calculations for the optimal assignment of the supply resources to the capacity targets.

Once the gaps and gluts for all skills j have been determined by solving (12) or (13), the last step of our end-to-end performance management process considers the actions that should be taken to address these gaps and gluts through hiring, training and retention. One approach to address this problem is based on a variation of our risk-based capacity planning models and optimization of Section 3. In this case, the objective (10) is modified to include the costs for hiring additional capacity possessing skill j, training existing capacity to acquire skill j, and retaining existing capacity with skill j, where existing capacity takes expected attrition into account (some of which can be incentivized for retention). The capacity vector $\mathbf{C}$ used in (10) and in the stationary loss network is also modified to reflect the sum of existing and retained capacity for skill j and new capacity for skill j based on hiring and training. Formally, the objective function for a main formulation of interest is given by

$$\begin{aligned}
\max_{\mathbf{C}} \quad & \sum_{r=1}^{|\mathcal{R}|} u_r \left(1 - L_r\right) \nu_r \;\; - \;\; \sum_{j=1}^{|\mathcal{L}|} v_j \, C_j \;\; - \;\; \sum_{j=1}^{|\mathcal{L}|} v^H_j \, C^H_j\\
& - \;\; \sum_{j=1}^{|\mathcal{L}|} v^T_j \, C^T_j \;\; - \;\; \sum_{j=1}^{|\mathcal{L}|} v^R_j \, C^R_j, && (14)
\end{aligned}$$

where $C_j = C^E_j + C^H_j + C^T_j + C^R_j$, v^H_j is the cost rate for hiring skill j capacity, v^T_j is the cost rate for training skill j capacity, v^R_j is the cost rate for retaining skill j capacity, and C^E_j, C^H_j, C^T_j and C^R_j is the amount of capacity existing, hired, trained and retained for skill j, respectively. The constraints of the optimization problem can also include a hiring cost budget B^H, a training cost budget B^T and a retention cost budget B^R, or more formally

$$\sum_{j=1}^{|\mathcal{L}|} v^H_j C^H_j \leq B^H, \quad \sum_{j=1}^{|\mathcal{L}|} v^T_j C^T_j \leq B^T, \quad \sum_{j=1}^{|\mathcal{L}|} v^R_j C^R_j \leq B^R.$$

Alternative formulations of interest include maximizing revenue or minimizing cost subject to constraints on the gross profit margin.

6. CASE STUDY

Our proposed solutions have been implemented as an integrated suite of tools that are currently used around the world by several IBM services, finance and HR organizations to support project delivery, workforce and engagement planning, sales-delivery interlock and higher level strategic planning. Although the final outputs of our system are optimal resource assignments and gap/glut reports, distinct components of the system have also been indepen-

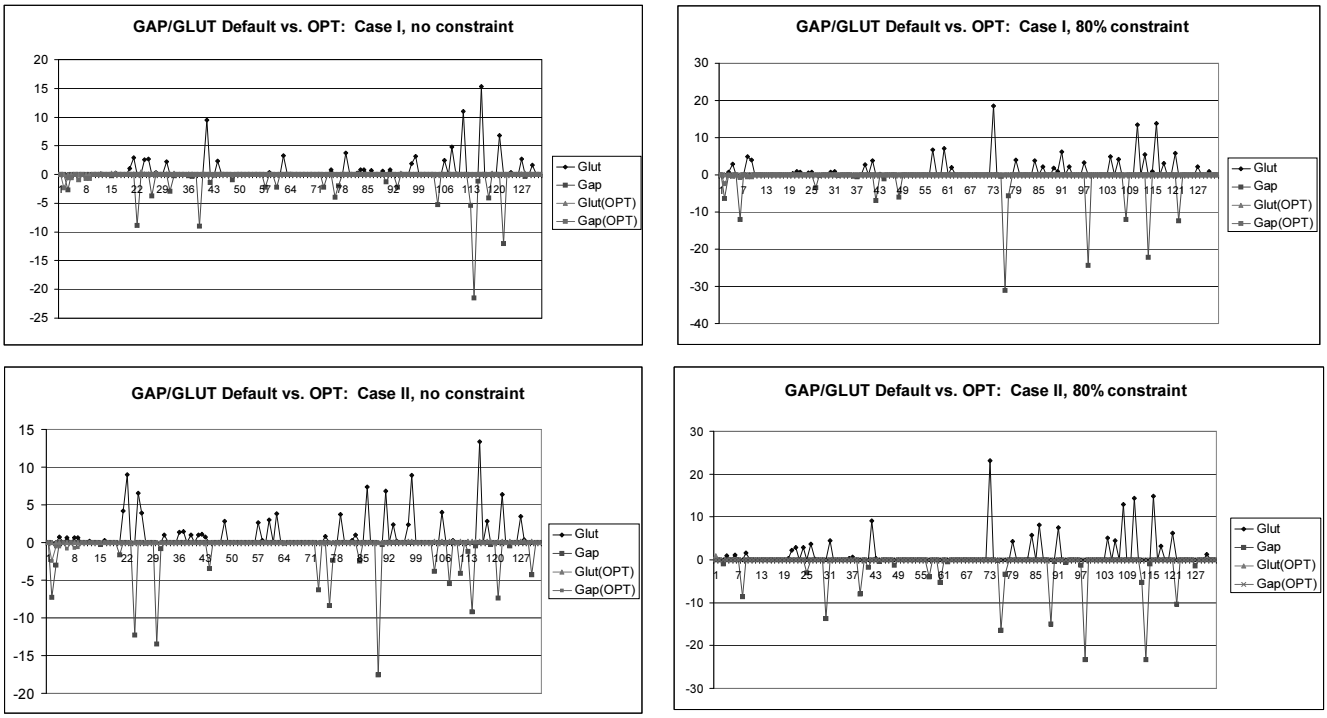

Figure 1: Comparison of Gap/Glut Results from Case Study

dently used in practice by different stakeholders to support additional business processes and various planning activities.

Given the nature of the data inputs and reports produced, our system serves as a trusted data source and fact base that integrates information from multiple systems and creates reports for multiple business functions. For example, in addition to computing the demand statement, the demand forecasting capability provides user interfaces for sales executives to follow the progression of the demand throughout the quarter, understand the organizational capability to deliver on business targets, and support the creation of sales and marketing initiatives. Similarly, the risk-based capacity planning capability, implemented as a decision-support aid for delivery executives, allows for studying the business performance and financial impacts of different delivery models and strategies, as well as different HR policies and actions. The supply forecasting capability is also implemented within a decision-support tool for talent management and HR organizations, allowing users to analyze historical workforce trends and dynamics, perform predictive modeling of future workforce dynamics, model future scenarios based on what-if adjustments to understand the effect of different actions/policies on workforce trends/dynamics, and optimize strategic decisions relative to business goals.

6.1 Deployment Experience and Observations

The management system provides rolling quarterly demand forecasts over a one year time horizon, where the demand coming from opportunities and expected deals becomes more dominant the further out into the future, as expected. Ongoing engagements typically account for around 80% of the demand in the immediate next quarter, but this decreases to around 60% in the second quarter and drops to 20-30% by the fourth future quarter. In contrast, demand from sales opportunities (already registered in the pipeline) tends to remain steady at around 20% for all future quarters, whereas the demand from expected deals typically rises from less than 5% in the immediate next quarter to 50-60% in the fourth future quarter.

Since the management system has been rolled out for a relatively short period of time, and since both the record tracking system and business processes are still evolving, there remains some limitations in our ability to make detailed comparisons between forecasted demand and actual deployment. As a result, we have only been able to carry out somewhat limited comparisons for some components of the system. In these comparisons, it is important to keep in mind that the forecasting results rely heavily on numerous human inputs provided through the normal course of business, such as the sales stage and the expected size of each opportunity, the expected remaining duration and resource requirements for an existing engagement, and the revenue targets. Thus the final accuracy of the system reflects not just the strength of the models, but also how rigorously various business information is provided and maintained in the system. It is expected that, as the system continues to be deployed and becomes an integral part of the services delivery business, more rigorous processes will be put in place to improve the accuracy of the input information, which in turn will further improve and refine our mathematical models and methods over time on a continual basis through a feedback control loop.

Our experience to date with deploying the system has been very positive. For pipeline forecasting, we have observed a typical accuracy range of 85-90% for total revenue using our statistical model, which represents an error reduction of more than 100% over forecasts computed using win probabilities manually estimated by the sales representatives themselves. The accuracy of the overall resource demand in terms of total hours has been observed to be generally higher than 90% when the accuracy of the revenue target is also higher than 90%, since this forecast is directly influenced by the revenue target. Our experience with risk-based capacity planning includes its effective use by executives to demonstrate and

evaluate the expected costs required to have enough resources to satisfy all forecasted demand with high probability (low risk), as well as to determine the per-skill capacity targets required to drive revenues while maintaining acceptable levels of costs and risks. We have observed the accuracy of our supply forecasting capability to be within a few percentage points for estimating the per-skill resource populations of organizations on the order of thousands or larger over 6-month to 1-year time horizons. The reductions in per-skill gaps and gluts obtained from our multi-skill gap/glut optimization has ranged from around 10-80% and 30-150%, respectively, in comparison with a default assignment using estimates on how supply will likely be assigned based on historical data and trends.

6.2 Numerical Experiments

We next present a representative sample of numerical results from the deployment of our end-to-end performance management solution with real data from the IT services industry, focusing on an example comprised of 110 service products and 132 skills. Our demand forecasting capability was used to construct two forecasts for the expected number of delivery engagements for each service product under different economic assumptions, yielding arrival rate vectors $\boldsymbol{\nu}^{D_1}$ and $\boldsymbol{\nu}^{D_2}$. This capability also provided the capacity requirement matrix $\mathbf{A}$ based on staffing templates determined for each service product. Our risk-based capacity planning capability was then applied to these inputs together with revenue and cost rates to obtain, for each demand forecast, two capacity planning target vectors $\mathbf{C}^O$ and $\mathbf{C}^C$ that maximize profit under no L_r constraint and a constraint of $L_r \leq \beta_r = 0.20$. Although the problem size is fairly large, our stochastic capacity planning solutions required less than a few seconds to compute using an interior-point NLP package (http://www.coin-or.org/). Our supply forecasting capability was used to determine the population of available resources and their skill mix. This together with each set of capacity planning targets were then fed as input to our multi-skill gap/glut capability to obtain the optimal assignment of supply resources to capacity targets and the corresponding gaps and gluts for all skills.

Figure 1 summarizes the final numerical results in comparison with those from an approach based on classical manufacturing supply chains. In particular, the 132-skill gaps/gluts are plotted for the four scenarios (two demand forecasts & two capacity targets) under the optimal assignment of multi-skill supply resources to per-skill capacity targets (OPT) and a basic default assignment. The basic default assignment refers to a policy that would be employed according to standard supply chain methods in the absence of our multi-skill gap/glut capability based on historical trends and data on how supply will likely be assigned to demand. We observe from these results that the optimal gaps and gluts are significantly lower than those under the default assignment. More specifically, the sum of the default gaps and gluts across all scenarios are in the range 200–400 whereas the same sum is reduced to be within the range 8–12 under our multi-skill gap/glut optimization; recall that these sums are taken over all 132 skills. Upon inputting these per-skill capacities from both the optimal and default assignments to our risk-based capacity planning models, we can further estimate the financial impact of these assignments which demonstrates improved profit under our optimal assignments in the range 350%–520% with the same costs and corresponding improvements in revenue.

7. CONCLUSIONS

Workforce management is becoming one of the most important factors in the ability of an organization to deliver products, grow revenue, be more profitable, and embrace the challenges of global integration. This is especially true for service-oriented businesses, and forward-thinking companies are investing in advanced mathematical models and methods to address performance management problems and realize a major competitive differentiator in the marketplace. In this paper, we presented a set of mathematical models and methods comprising an end-to-end performance management solution for the delivery of IT services, together with a case study applying our solution to real data from the IT services industry.

8. REFERENCES

[1] AMR. Workforce management landscape: The right people in the right place at the right time. AMR Res. Rep., 2006.

[2] J. Anselmi, Y. Lu, M. Sharma, M. Squillante. Improved approximations for the Erlang loss model. *QUESTA*, 63:217–239, 2009.

[3] M. Bazaraa, H. Sherali, C. Shetty. *Nonlinear Programming: Theory and Algorithms*. Wiley, 1993.

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

[5] S. Bhadra, Y. Lu, M. Squillante. Optimal capacity planning in stochastic loss networks with time-varying workloads. In *ACM SIGMETRICS*, pp. 227–238, 2007.

[6] B. Chen, R. Tai, R. Harrison, Y. Pan. Novel hybrid hierarchical-k-means clustering method (h-k-means) for microarray analysis. In *IEEE BCSBW*, 2005.

[7] J. Hu, B. Ray, M. Singh. Statistical methods for automated generation of service engagement staffing plans. *IBM J. Res. Dev.*, (3), 2007.

[8] J. Hu, M. Singh, A. Mojsilovic. Categorization using semi-supervised clustering. In *ICPR*, 2008.

[9] J. Hu, M. Singh, A. Mojsilovic. Using data mining for accurate resource and skill demand forecasting in services engagements. In *KDD Workshop on Data Mining for Business Applications*, 2008.

[10] P. Hunt, F. Kelly. On critically loaded loss networks. *Adv. App. Prob.*, 21(4):831–841, 1989.

[11] K. Jung, Y. Lu, D. Shah, M. Sharma, M. Squillante. Revisiting stochastic loss networks: Structures and algorithms. In *ACM SIGMETRICS*, pp. 407–418, 2008.

[12] K. Jung, Y. Lu, D. Shah, M. Sharma, M. Squillante. Revisiting stochastic loss networks: Structures and algorithms. Submitted, 2009.

[13] F. Kelly. Loss networks. *Ann. App. Prob.*, 1(3):319–378, 1991.

[14] G. Louth, M. Mitzenmacher, F. Kelly. Computational complexity of loss networks. *Theor. Comp. Sci.*, 125(1):45–59, 1994.

[15] Y. Lu, A. Radovanović, M. Squillante. Optimal capacity planning in stochastic loss networks. *ACM PER*, 35(2), 2007.

[16] Y. Lu, A. Radovanović, M. Squillante. Optimal capacity planning in general stochastic loss networks. Preprint, 2009.

[17] Y. Lu, M. Sharma, M. Squillante. Stochastic analytics and optimization for workforce evolution. IBM Res. Rep., 2006.

[18] P. Momcilovic, M. Squillante. On throughput in linear wireless networks. In *ACM Symposium on Mobile Ad Hoc Networking and Computing*, 2008.

[19] W. Whitt. Blocking when service is required from several facilities simultaneously. *AT&T Bell Lab. Tech. J.*, 64(8):1807–1856, 1985.