Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds.

A TWO-PHASE APPROACH FOR STOCHASTIC OPTIMIZATION OF COMPLEX BUSINESS PROCESSES

Soumyadip Ghosh
Aliza R. Heching
Mark S. Squillante

Business Analytics and Mathematical Sciences
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598 USA

ABSTRACT

Business process modeling is a well established methodology for analyzing and optimizing complex processes. To address critical challenges in ubiquitous black-box approaches, we develop a two-stage business process optimization framework. The first stage is based on an analytical approach that exploits structural properties of the underlying stochastic network and renders a near-optimal solution. Starting from this candidate solution, the second stage employs advanced simulation optimization to locally search for optimal business process solutions. Numerical experiments demonstrate the efficacy of our approach.

1 INTRODUCTION

Business processes, the defined sets of activities by which organizations set out to achieve specific business goals, are often highly complex due to their associated large number of interrelated decisions and high degree of uncertainty. Business process modeling, the process by which the structural representation of the business process is created and the relationship between different entities in the business process is defined, has been leveraged as an effective tool to improve and optimize business process performance and/or financial metrics, thus contributing to improvements in these overall metrics for an organization. Representative examples of business processes include order management by an online retailer and the flow of patients through a hospital emergency room. A critical component associated with the successful execution of many business processes is optimal resource capacity management. The resources in business processes may be human resources (e.g., physicians treating patients in a hospital emergency room) or they may be machine resources (e.g., resource capacity allocation in cloud computing environments).

Complex business processes are often characterized by the existence of one or more types of resources serving one or more system demand flows where types of resources may be differentiated according to their attributes (e.g., skills or service rates) and demand flows may be differentiated according to their workloads (e.g., arrival or processing rates) and quality of service (e.g., performance guarantees). The objective of resource capacity management is to determine the capacity allocation for each type of resource supporting the business process that optimizes the organizational performance/financial goals. Often, constraints are imposed that may limit the value of the achievable performance/financial goal or the total quantity of resources that may be allocated. Contributing to the complexity of this resource capacity management problem is the uncertainty inherent in business processes. This uncertainty may be present, for example, due to uncertain demand arrival patterns or uncertain processing requirements.

978-1-4799-2076-1/13/$31.00 ©2013 IEEE

1.1 Common Solution Approach

Stochastic networks are often used to model the topology of a series of activities that comprise the business process and to represent the dynamics and uncertainty associated with the resource capacity management optimization problem. To facilitate optimal decision-making in the business process, the stochastic network is analyzed. Optimal solutions yielded from the analysis of the stochastic network are translated into requirements for optimal management of the business process. Two approaches are commonly adopted for analyzing and deriving optimal solutions for stochastic networks associated with capacity management problems for complex business environments. The first approach relies on the application of product-form stochastic network results. Although the underlying stochastic network typically does not satisfy the strong restrictions of a product-form queueing network, the results of this class of queueing networks are used to provide simple approximate solutions of the resource allocation problem. The solution is typically combined with additional heuristics to yield a final solution (see, e.g., Menasce and Almeida 2000). Although these solutions can be quickly obtained, the approximations required by this approach often result in solutions that suffer from inaccuracies that are of both theoretical and practical concern.

The second approach that is commonly adopted is based upon simulation-based optimization. Unlike the simplifications that are often required when adopting analytical approaches to solve stochastic network problems, simulation has the benefit of allowing for the incorporation of all forms of uncertainty as well as complex interactions that may be present in the stochastic network. Within this solution approach, the literature may be classified into two broad categories. The first category of approaches applies metaheuristics such as tabu and scatter search to control the sequence of simulation runs that are evaluated in order to identify the optimal solution to the resource capacity management problem (see, e.g., Nelson and Henderson 2007). At each step, the simulated objective function value for the current set of decision variables is compared with the objective function values yielded for previously evaluated sets of decision variables. The "best-to-date" set of decision variables (i.e., best objective function values) is noted. A stopping criterion (e.g., run time-based or percentage improvement) is used to determine when the procedure should halt and the "best-to-date" set of decision variables is declared the optimal solution. Although metaheuristics have been incorporated in a wide array of simulation software products that support optimization (e.g., Crystal Ball, Arena, AnyLogic, and SIMUL8), they have associated accuracy and long runtime concerns for business processes because they fail to consider any structure of the underlying stochastic network.

The second category of simulation-based optimization approaches consists of direct methods, such as stochastic approximation, which analyze the convergence behavior of the objective function to guide the selection of simulation runs to be evaluated (see, e.g., Asmussen and Glynn 2007). The use of direct methods such as stochastic approximation is not as prevalent as the use of metaheuristics. A challenge in the efficient application of stochastic approximation is the need to identify "good" starting values for key parameters. However, these "good" starting values vary depending upon the problem instance being solved, and thus limit the successful application of stochastic approximation in practice. Direct methods may also suffer from long runtimes in large part because of the numerous parameters involved in each method that must be set via experimental tweaking for every problem instance.

While simulation approaches permit an exact representation of the stochastic network without simplifying approximations, significant drawbacks are the temporal and computational requirements to identify an optimal solution for large-scale business processes. Indeed, real-world problems involve multiple resource classes and demand classes, leading to a multidimensional stochastic network that increases the time required to identify an optimal solution. A recent study illustrates how simulation-based optimization may require on the order of days to determine optimal resource capacity levels in a class of business processes (Heching and Squillante 2013). As the complexity and scale of business processes continue to grow, there remains a critical need to address the costs in both time and resources of a purely simulation-based optimization approach even with continuing improvements in simulation-based optimization methods.

1.2 Our Solution Approach

In this paper, we describe a two-stage solution approach that has been successfully applied to solve resource allocation problems across diverse complex business processes; see Dieker et al. (2012) and Heching and Squillante (2013). Our solution combines the efficiency of an analytical approach with the accuracy of a simulation approach. The first stage of our two-stage approach exploits mathematical approximations to efficiently obtain a nearly optimal solution to the business process optimization problem. Depending on the nature of the complex business process and the underlying stochastic network, different mathematical methods may be adopted. Further refinement of the solution may be required to address inaccuracies due to the approximations and relaxations required to efficiently obtain analytical approximations to stochastic network representations of complex business processes. Hence, the approximate solution from the first stage is used as a starting point for the second stage of our solution approach. In this second stage, simulation-based optimization is used to identify a high-quality optimal solution to the original stochastic network, where any simulation method may be employed including metaheuristics and direct methods.

As illustrated in Fig. 1, the first stage of our general solution approach can be further divided into two phases. The first phase is based on direct stochastic approximations and consists of applying mathematical methods to simplified approximate models, parameterized by the original stochastic network inputs, in order to yield an analogous resource allocation problem that is amenable to direct analysis without further aid of any simulation. The second phase is based on stochastic decoposition and functional-form approximations, and consists of a combination of mathematical approximations and numerical or simulation-based solvers, using as a starting point the first-phase results. More specifically, the objective function of the original stochastic network is represented using parametrized canonical functional forms that arise from the stochastic network literature, where some parameters of the functional forms can only be estimated via numerical or simulation solvers; the optimal solution is then derived in terms of the separable functional form and an iterative procedure is used together with the outputs of numerical or simulation solvers to obtain a very good candidate solution for the second stage of our methodology. Depending upon the application model of interest, one may either use this combination of both phases of the first stage of our general solution approach, or simply choose one of these phases to provide the first-stage results. Finally, the resource allocation decisions from the first stage subsequently serve as a starting point for the second stage of our general solution approach. The second stage consists of a general search capability based on simulation-based optimization methods that deal directly with the original stochastic network to further improve upon the first-phase results and obtain a locally optimal solution for the original business process.

There are several important reasons for adopting such a two-stage solution approach to support the stochastic optimization of complex business processes. It is of primary importance to be able to determine optimal solutions in a highly accurate and efficient manner. The combination of the two phases of the first

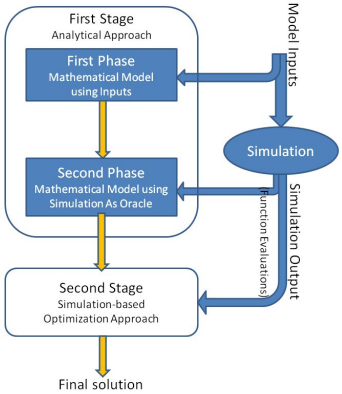

Figure 1: Two-Stage Solution Framework

stage yields nearly optimal solutions within only a few numerical/simulation runs, and thus significantly outperforms state-of-the-art approaches for evaluating and optimizing business process objective functions by several orders of magnitude in reduced time and resources. Such tremendous improvements realized through our first-stage methodology for stochastic optimization of complex processes enables a much broader and deeper exploration of the entire business process design space. When this nearly optimal solution is used as a starting point in the second-stage methodology, the latter may be exploited more surgically to explore regions of the design space of greatest importance or sensitivity to obtain optimal solutions. This in turn provides significant improvements in both the efficiency and quality of the stochastic optimization of large complex processes in practice as compared with state-of-the-art approaches, thereby successfully benefiting from the positive aspects of its analytical-based and simulation-based components.

The remainder of this paper is organized as follows. Sections 2 and 3 respectively detail the first and second phases of our first-stage methodology, including analytical approaches for the analysis and optimization of a general class of stochastic models that yield a nearly optimal solution. Section 4 briefly discusses our second-stage methodology, which renders the final solution to the original stochastic optimization problem. Section 5 provides some examples where this approach is used, along with representative numerical results.

2 FIRST PHASE OF STAGE 1 METHODOLOGY

The first phase of our stage 1 methodology consists of developing an approximate mathematical model of the original multidimensional stochastic process that underlies the complex business process, deriving a mathematical analysis of this stochastic process approximation, and then deriving an optimal solution to the corresponding stochastic optimization problem. Such a mathematical analysis renders an approximate expression for the objective function of the original stochastic optimization problem, often in closed form, which can be solved more easily than the original optimization problem. We then determine an optimal solution to the resulting approximate stochastic optimization problem, which often represents a nearly optimal solution to the original problem. These results serve as a starting point for the second phase of our stage 1 methodology, though in some cases our first-phase results are sufficiently accurate to directly serve as a starting point for our stage 2 methodology.

Many stochastic process approximations are available to us. One set of general approaches is based on stochastic decomposition of the complex multidimensional stochastic process into a combination of various forms of simpler processes with reduced dimensionality (see, e.g., Squillante 2011), including nearly completely decomposable stochastic systems, asymptotic independence together with fixed-point equations, and priority structural properties together with recursion. Another set of general approaches is based on mathematical analysis of the complex stochastic process and associated control problem in some limiting regime (see, e.g., Chen and Yao 2001), including fluid limits, diffusion limits and strong approximations respectively characterizing the asymptotic behavior via a functional strong law of large numbers, a functional central limit theorem and a functional strong approximation theorem. In general, any of the available algorithms and methods from applied probability, stochastic optimization/control, and mathematical programming can be exploited as part of the first phase of our stage 1 methodology.

We consider two specific instances of our first-phase methodology. The first of these instances, developed in Dieker et al. (2012), concerns a general feedforward stochastic network that is initially approximated by the closest instance (Dieker et al. 2008) of a product-form Brownian network (Harrison and Williams 1992). This yields a closed-form expression for the performance measures of interest and then the optimal resource capacities are derived for the corresponding stochastic optimization problem, which are subsequently used as a starting point for a specific instance of the second phase of our stage 1 methodology (see Section 3). Another instance of our first-phase methodology, developed in Heching and Squillante (2013), concerns a general multi-class stochastic network that is initially approximated by a collection of simplified stochastic models. Then large deviations, strong approximation and other results are derived for the approximate stochastic network. These results are used in turn to derive an efficient solution to the corresponding stochastic optimization problem based on a combination of mathematical programming algorithms and

the approximate stochastic analysis. The corresponding first-phase results are sufficiently accurate and complete to directly serve as a starting point for our stage 2 methodology (see Section 4). In what follows, we provide a summary of some of the technical details from this second instance of the first phase of our stage 1 methodology; refer to Heching and Squillante (2013) for additional details.

We begin by defining and analyzing an approximate stochastic model. We partition the time horizon of interest over time-varying workloads and resource capacity work shifts into stationary intervals and then define a stochastic model for each stationary interval such that these stationary models are combined to represent the entire time horizon. For each stationary interval indexed by $i \in \mathbb{I}$, the model involves a set $\mathbb{J}$ of priority queueing systems, one for each group of resources indexed by $j \in \mathbb{J}$. A customer request belongs to one of a set $\mathbb{K}$ of request classes, indexed by $k \in \mathbb{K}$. Requests of class k arrive to the queueing system according to a stochastic process $\{A_{i,k}(t); t \geq 0\}$ with $A_{i,k}(t) := \sup\{n : \mathscr{A}_{i,k}(n) \leq t\}$ and finite rate $\lambda_{i,k}$, where $\mathscr{A}_{i,k}(n) := \sum_{m=1}^{n} a_{i,k}^{(m)}$, $\mathscr{A}_{i,k}(0) := 0$, the random variable (r.v.) $a_{i,k}^{(n)}$ is the interarrival time between the $(n-1)$st and nth class k requests, and $a_{i,k}^{(0)} := 0$, $n \geq 1$. Upon arrival, a class k request is routed to the queueing system of group j in interval i according to a routing policy that renders a corresponding class-group stochastic arrival process $\{A_{i,j,k}(t); t \geq 0\}$ with $A_{i,j,k}(t) := \sup\{n : \mathscr{A}_{i,j,k}(n) \leq t\}$ and finite rate $\lambda_{i,j,k}$, where $\mathscr{A}_{i,j,k}(n) := \sum_{m=1}^{n} a_{i,j,k}^{(m)}$, $\mathscr{A}_{i,j,k}(0) := 0$, the r.v. $a_{i,j,k}^{(n)}$ is the interarrival time between the $(n-1)$st and nth class k requests served by group j in interval i, and $a_{i,j,k}^{(0)} := 0$, $n \geq 1$. The class-group sequences of interarrival time r.v.s $a_{i,j,k}^{(n)}$ are such that $\lambda_{i,k} = \sum_{j \in \mathbb{J}} \lambda_{i,j,k}$ with $\lambda_{i,j,k}$ fixed to be 0 whenever group j does not have the appropriate capabilities to serve class k requests. The times required for resource group j to serve class k requests in interval i are governed by a stochastic process $\{S_{i,j,k}(t); t \geq 0\}$ with $S_{i,j,k}(t) := \sup\{n : \mathscr{S}_{i,j,k}(n) \leq t\}$ and finite rate $\mu_{i,j,k}$, where $\mathscr{S}_{i,j,k}(n) := \sum_{m=1}^{n} s_{i,j,k}^{(m)}$, $\mathscr{S}_{i,j,k}(0) := 0$, and the r.v. $s_{i,j,k}^{(n)}$ is the time required to serve the nth class k request by group j in interval i, $n \geq 1$.

The queueing system for each group employs a fixed-priority scheduling policy; requests of class k are given priority over requests of class k' for all $k < k'$ with $k, k' \in \mathbb{K}$. A preemptive-resume scheduling discipline is deployed across request classes in which the serving of preempted requests is resumed from the point where they left off without any overhead. Requests within each class are served in a first-come, first-served manner. Let $C_{i,j}$ denote the number of resources (capacity) that comprises group j, which we relax by interpreting $C_{i,j} \in \mathbb{R}_+$ to be a capacity scaling variable for the processing rate of a corresponding multiclass $GI/GI/1$ fixed-priority queueing system for each group j and stationary interval i. Another relaxation of the original stochastic model concerns the class-group routing decision process $\Lambda_{i,j,k}(\cdot)$ for each request class k, resource group j and stationary interval i. To simplify the stochastic analysis, we assume these routing decisions to be probabilistic such that a class k request is routed to group j with probability $P_{i,j,k}$, independent of all else. We therefore have $\Lambda_{i,j,k} = P_{i,j,k}\lambda_{i,k}$, and thus determining the optimal routing decision variables reduces to obtaining the optimal routing probabilities $P_{i,j,k}^{*}$.

Performance measures of interest include the sojourn time process $\mathbf{T}_{i,j,k} = \{\mathbf{T}_{i,j,k}(t); t \geq 0\}$, the aggregated workload process $\mathbf{Z}_{i,j,k} = \{\mathbf{Z}_{i,j,k}(t); t \geq 0\}$ and the cumulative idle time process $\mathbf{Y}_{i,j,k} = \{\mathbf{Y}_{i,j,k}(t); t \geq 0\}$, where $\mathbf{T}_{i,j,k}(t)$ denotes the total sojourn time of class k requests at the group j queueing system in stationary interval i at time t, $\mathbf{Z}_{i,j,k}(t)$ denotes the total amount of existing work at the group j queueing system in stationary interval i comprised of requests in classes 1 through k that are either in queue or in service at time t, and $\mathbf{Y}_{i,j,k}(t)$ denotes the cumulative amount of time that the group j queueing system in stationary interval i does not serve requests in classes 1 through k during $[0,t]$. Define the net-put process $\mathbf{N}_{i,j,k}(t)$ to be the total workload input from request classes 1 through k at the group j queueing system in stationary interval i during $[0,t]$ minus the work that would have been completely served if the queueing system was never idle, $\mathbf{U}_{i,j,k}(t)$ to be the total amount of time spent serving class k requests at the group j queueing system in stationary interval i during $[0,t]$, $\mathbf{V}_{i,j,k}(t)$ to be the time that a class k request would spend at the group j queueing system in stationary interval i if it arrived at time t, and $\mathbf{G}_{i,j,k}(t)$ to be the time at which the first class k request arrives to the group j queueing system in stationary interval i

during $[t,\infty)$. Let $\mathscr{M}_X(\theta) = \mathbb{E}[e^{\theta X}]$ denote the moment generating function of a r.v. X, and let $f(n) \sim g(n)$ denote that $\lim_{n\to\infty} f(n)/g(n) = 1$ for any functions f and g. Further define $\rho^+_{i,j,k} := \sum_{k' \le k, k' \in \mathbb{K}} \rho_{i,j,k'}$, $\rho_{i,j,k} := \lambda_{i,j,k}/(C_{i,j}\mu_{i,j,k})$, and $(x)^+ := \max\{x, 0\}$.

Now, we turn to consider the derivation of results for key performance measures, focusing on, as a representative illustration, deriving strong approximations of the class k sojourn time processes for the above relaxation of the original stochastic model. Using standard definitions and notation for r-strong continuous functions/processes and strong approximations over the space $\mathscr{D}^{|\mathbb{K}|}$ of $|\mathbb{K}|$-dimensional real-valued functions on $[0,\infty)$ that are right-continuous with left limits (Chen and Yao 2001), we denote by $\tilde{\mathbf{X}} = \{\tilde{\mathbf{X}}(t); t \ge 0\}$, $\tilde{\mathbf{X}}(t) = mt + \hat{\mathbf{X}}(t)$, a strong approximation of a stochastic process $\mathbf{X} = \{\mathbf{X}(t); t \ge 0\}$ in $\mathscr{D}^{|\mathbb{K}|}$ under appropriate conditions. In such cases, we write $\mathbf{X}(t) \overset{r}{\approx} \tilde{\mathbf{X}}(t)$ or $\mathbf{X} \overset{r}{\approx} \tilde{\mathbf{X}}$. Then, for the previously defined performance measures of interest, we have:

$$\mathbf{Z}_{i,j,k}(t) = \mathbf{N}_{i,j,k}(t) + \mathbf{Y}_{i,j,k}(t), \quad \mathbf{T}_{i,j,k}(t) = \mathbf{V}_{i,j,k}(\mathbf{G}_{i,j,k}(t)), \quad 0 \le \mathbf{G}_{i,j,k}(t) - t \le a_{i,j,k}(A_{i,j,k}(t)+1), \quad (1)$$

$$\mathbf{N}_{i,j,k}(t) = \sum_{k'=1}^{k} \mathscr{S}_{i,j,k'}(A_{i,j,k'}(t)) - t, \qquad \mathbf{Y}_{i,j,k}(t) = t - \sum_{k'=1}^{k} \mathbf{U}_{i,j,k'}(t) = \sup_{0\le s\le t}\{-\mathbf{N}_{i,j,k}(s)\}, \quad (2)$$

$$\mathbf{V}_{i,j,k}(t) = \mathbf{Z}_{i,j,k}(t) + \sum_{k'=1}^{k-1} [\mathscr{S}_{i,j,k'}(A_{i,j,k'}(\mathbf{V}_{i,j,k}(t)+t)) - \mathscr{S}_{i,j,k'}(A_{i,j,k'}(t))] + \mathscr{S}_{i,j,k}(A_{i,j,k}(t)) - \mathscr{S}_{i,j,k}(A_{i,j,k}(t)-1). \quad (3)$$

In Theorem 1 we establish the desired strong approximations for the key performance measures of the sojourn time and aggregated workload processes. We subsequently leverage these results to directly obtain the first moment, second moment, or tail probability associated with the sojourn time distribution for class k requests served by group j in stationary interval i. Refer to (Heching and Squillante 2013).

Theorem 1 For class $k \in \mathbb{K}$ requests served at the $GI/GI/1$ fixed-priority queueing system of resource group $j \in \mathbb{J}$ with capacity $C_{i,j}$ in stationary interval $i \in \mathbb{I}$ such that the strong approximation assumptions hold for some $r \in (2,4)$, we have $(\mathbf{Z}_{i,j,k}, \mathbf{T}_{i,j,k}) \overset{r}{\approx} (\tilde{\mathbf{Z}}_{i,j,k}, \tilde{\mathbf{T}}_{i,j,k})$, where

$$\tilde{\mathbf{Z}}_{i,j,k}(t) = \tilde{\mathbf{N}}_{i,j,k}(t) + \tilde{\mathbf{Y}}_{i,j,k}(t), \quad (4)$$

$$\tilde{\mathbf{T}}_{i,j,k}(t) = \frac{\tilde{\mathbf{Z}}_{i,j,k}(t) + R_{i,j,k}}{1-\rho^+_{k-1}} \ge \frac{\tilde{\mathbf{Z}}_{i,j,k}(t) + s_{i,j,k}}{1-\rho^+_{k-1}}, \quad (5)$$

$$\tilde{\mathbf{N}}_{i,j,k}(t) = (\rho^+_{i,j,k} - 1)t + \sum_{k'=1}^{k} [\hat{A}_{i,j,k'}(t)/\mu_{i,j,k'} + \hat{\mathscr{S}}_{i,j,k'}(\lambda_{i,j,k'}t)], \quad (6)$$

$\tilde{\mathbf{Y}}_{i,j,k}(t) = \sup_{0\le s\le t}\{-\tilde{\mathbf{N}}_{i,j,k}(s)\}^+$, and $\tilde{\mathbf{Z}}_{i,j,k}(t)$, $\tilde{\mathbf{T}}_{i,j,k}(t)$ are r-strong continuous.

Next, we consider the formulation of a corresponding stochastic optimization problem whose objective is to determine the capacity of resource groups j and the routing of class k requests to resource groups j that maximize profit in expectation under the foregoing stochastic model across all stationary intervals comprising the time horizon together with the overlap of work shifts, subject to model inputs and constraints. We therefore seek to determine the optimal capacity $C^*_{i,j}$ for resource groups j and the optimal routing decision process $\Lambda^*_{i,j,k}(\cdot)$ for class k requests and groups j, both in stationary intervals i, over the time horizon of interest $\mathscr{T}$ with respect to every performance guarantee, given class-group routing constraints and exogenous arrival and service time processes $A_{i,k}(\cdot)$ and $S_{i,j,k}(\cdot)$. The optimal class-group routing decision vector process $\{\mathbf{\Lambda}^*(t); t \ge 0\}$ with $\mathbf{\Lambda}^*(t) := (\Lambda^*_{i,j,k}(t))$ includes determining the set of stochastic arrival processes $A^*_{i,j,k}(t)$ with finite rates $\Lambda^*_{i,j,k} = \lambda^*_{i,j,k}$. To simplify the formulation, suppose the optimal routing of class k requests satisfy $\lambda_{i,j,k}/(C_{i,j}\mu_{i,j,k}) < 1$. Further, consider the performance guarantees of

class k to be such that penalties are incurred with respect to $\left(\mathbb{P}[T_{i,j,k} > \mathsf{Z}_k] - \alpha_k\right)^+$, where $T_{i,j,k}$ denotes the corresponding generic stationary sojourn time r.v. The revenue, penalty, and cost functions are linear in the total number of requests, number of performance guarantee violations, and number of resources.

Define the capacity vector $\mathbf{C} := (C_{i,j})$ and the class-group routing rate vector $\mathbf{\Lambda} := (\Lambda_{i,j,k})$. We then have the following general formulation of the stochastic optimization problem over the time horizon $\mathscr{T}$:

$$\begin{aligned}
\max_{\mathbf{C},\mathbf{\Lambda}} \quad & \sum_{i\in\mathbb{I}}\sum_{j\in\mathbb{J}}\sum_{k\in\mathbb{K}} \mathbb{E}[\mathscr{N}_{i,j,k}(\mathscr{T})]\mathscr{R}_{j,k} - \mathbb{E}[\mathscr{N}_{i,j,k}(\mathscr{T})]\mathscr{P}_{j,k}\left(\mathbb{P}[T_{i,j,k} > \mathsf{Z}_k] - \alpha_k\right)^+ - \mathscr{C}_{j,k}C_{i,j}, && (7)\\
\text{s.t.} \quad & \sum_{j\in\mathbb{J}} \Lambda_{i,j,k} = \lambda_{i,k}, && \forall i \in \mathbb{I}, \forall k \in \mathbb{K},\\
& \Lambda_{i,j,k} = 0, \quad \text{if } \mathscr{I}(j,k) = 0, && \forall i \in \mathbb{I}, \forall j \in \mathbb{J}, \forall k \in \mathbb{K},\\
& \Lambda_{i,j,k} \geq 0, \quad \text{if } \mathscr{I}(j,k) = 1, && \forall i \in \mathbb{I}, \forall j \in \mathbb{J}, \forall k \in \mathbb{K},
\end{aligned}$$

where $\mathscr{N}_{i,j,k}(t)$ is the cumulative number of class k requests routed to group j in interval i through time t, and $\mathscr{I}\{j,k\} := \mathbb{1}\{\text{class } k \text{ requests can be served by resource group } j\}$. The capacity vector $\mathbf{C}$ and routing vector process $\mathbf{\Lambda}(\cdot)$ are the decision variables we seek to obtain, with all other variables as input parameters.

Towards obtaining the optimal solution, we first determine the minimum resource capacities required to ensure that each of the resource group queueing systems is stable. This must hold for any feasible solution of the stochastic optimization problem and therefore requires that $\rho^+_{i,j,|K|} < 1$ for all $i \in \mathbb{I}$ and $j \in \mathbb{J}$. The above model relaxations allow us to focus on the capacity vector $\mathbf{C} \in \mathbb{R}_+^{|\mathbb{I}|\times|\mathbb{J}|}$ and the class-group routing rate vector $\mathbf{\Lambda} \in \mathbb{R}_+^{|\mathbb{I}|\times|\mathbb{J}|\times|\mathbb{K}|}$ as the decision variables of interest. The remaining step for an explicit formulation is to deal with the probabilities $\mathbb{P}[T_{i,j,k} > \mathsf{Z}_k]$ in (7). Based on our stochastic analysis above, we directly substitute (5) and derive an analytical expression for the objective function (7).

The best mathematical programming methods to solve the stochastic optimization problem will depend upon the fundamental properties of the specific optimization problem of interest. When the objective and constraint functions are concave in $(\mathbf{C},\mathbf{\Lambda})$, we exploit these and related properties of the equations in Theorem 1 together with convex programming methods (see, e.g., Boyd and Vandenberghe 2004) to efficiently obtain the unique optimal solution. More generally, when the objective or constraint functions are not concave, we exploit advanced nonlinear programming methods to efficiently obtain an (locally) optimal solution of the stochastic optimization problem by leveraging some of the best interior-point algorithms and implementations. We further note that this mathematical programming solution approach applies more generally to problem instances with nonlinear revenue, penalty and cost functions.

3 SECOND PHASE OF STAGE 1 METHODOLOGY

The second phase of our stage 1 methodology also works with an approximate mathematical model of the original stochastic process. However, the approach is distinguished from the first phase in its use of a numerical or simulation solver for further evaluations of the approximate model. Hence, the numerical or simulation solver serves as a function evaluation oracle. The reasons for further refinements of the first-phase results are due to the complexity of solving resource capacity management problems, which are due in turn to the technical difficulties of solving for functionals of general multidimensional stochastic processes involving various dependencies and dynamic interactions among the different dimensions. Often, the mathematical methods exploited in the first phase of our stage 1 methodology do not adequately and completely capture these complexities. In such cases, the second phase of our stage 1 methodology is invoked. We consider in this section one particular approach for the second phase of our stage 1 methodology that addresses such difficulties. Due to the highly nonlinear and possibly nonconvex nature of resource capacity management problems in general, our focus here is on finding "good" local optima.

The second-phase model renders an approximation for the objective function of the original problem based on information from the foregoing function evaluation oracle. Here, the functional evaluations are often localized to a current region of interest in the solution space, and thus the approximation of the true

objective function is good only locally. We then derive an optimal solution to the approximate objective function, which may also be constrained to fall within some region of trust around the current point of interest. The procedure may iteratively follow up with the re-construction of the approximate model using functional evaluations around this next candidate solution and further optimization of the new model. Hence, this second phase is usually based on a fixed-point iteration approach where observed function values at the current iterate determine the candidate resource capacity allocation for the next iterate. The next iterate will re-balance resource capacity allocation such as to obtain desirable changes in the value of the approximation to the true objective function of interest, and is thus consistent with optimizing the true objective function at least in approximation. This process repeats, forming the basis of an efficient fixed-point iteration that renders a nearly (locally) optimal solution to the resource capacity management problem. The iterations proceed until the candidate solutions identified exhibit little change under some appropriately defined stopping criterion. Depending on the stochastic network setting in which our general solution framework is applied, the required function evaluation may be obtained via (a combination of) advanced analytical (e.g., Harrison and Williams 1992), numerical (e.g., Dai and Harrison 1992) or simulation-based methods (e.g., Asmussen and Glynn 2007). As a result, this second phase may be applied to stochastic networks that are analytically intractable as long as they can be simulated or otherwise numerically evaluated.

The key distinction between this second phase approach of the first stage and the general simulation-based optimization of the second stage is the assumption of a (local) separable functional form for the objective function. For the second phase to substantially improve upon the objective function value beyond the first phase, the separable functional form must be carefully selected. Our primary example for this phase of the methodology is the problem of identifying optimal resource allocations in a stochastic network setting that minimizes a functional of the queue lengths at each node in the network. The general functional form for the queue length at a node of the network is not known as a closed-form function of the resource allocation at all the nodes of the network. In this example, our second-phase functional form for the queue length is assumed to be (locally) similar to the queue length of a $G/G/1$ queue with similar arrival and service characteristics. Our iterative algorithm then updates resource allocations based on the square root of the observed queue lengths, as motivated and formalized below. Roughly speaking, our updating rule is derived from an appropriate separable functional form for the performance metrics of each station in the network, such as expected steady-state queue length or expected steady-state sojourn time at the queue. The functional form is given by $\tau/(\beta-\lambda)$, where λ and β are the arrival and service rates for the queue and τ is a function of various characteristics of the arrival and service processes at all stations in the network, and must be estimated from evaluations of the true business process. This particular functional form naturally arises in all known queueing formulas.

In what follows, we provide a summary of some of the technical details from one specific instance of the second phase of our stage 1 methodology, developed in Dieker et al. 2012. We formalize our approach in a setting where the goal is to minimize the sum of the weighted expected queue lengths in a stochastic network serving customers of a single class and is subject to a budgetary constraint. The discussion is geared towards application of our approach to generalized Jackson networks (e.g., Chen and Yao 2001) and their Brownian counterparts (e.g., Harrison and Williams 1987). In particular, $\boldsymbol{\gamma}$ represents the effective arrival rate vector and $\boldsymbol{\beta}$ represents the vector of service rates. (Further parameters of the network, such as the routing matrix and the exact external interarrival and service distributions, need not be specified to present our approach and thus we do not introduce them here.) Let $Z_i^{\boldsymbol{\beta}}$ denote the steady-state queue length at the i-th station. (Alternatively, one can similarly study steady-state sojourn times). The dependence on $\boldsymbol{\beta}$ is made explicit since we are interested in comparing a functional of the steady-state vector $\boldsymbol{Z}^{\boldsymbol{\beta}}$ as we change the service-rate vector $\boldsymbol{\beta}$. Assume that each unit of resource capacity at station i costs c_i, comprising a cost vector $\boldsymbol{c}$, and that we have a total budget of C for allocating resources in the network.

The main stochastic optimization formulation is given by

$$\text{(OPT)} \qquad \min_{\beta\in(0,\infty)^L} \sum_{i=1}^{L} w_i \mathbb{E} Z_i^{\boldsymbol{\beta}} \qquad \text{s.t.} \quad \langle \mathbf{c}, \boldsymbol{\beta} \rangle \le \mathrm{C}, \quad \beta_{\mathrm{i}} > \gamma_{\mathrm{i}}, \quad \mathrm{i} = 1, \ldots, \mathrm{L}.$$

The expected steady-state queue lengths weighted by a vector $\boldsymbol{w}$ is minimized, subject to the constraints that (i) spend may not exceed the budget C and (ii) the queueing system is stable. Throughout, we shall assume $\langle \mathbf{c}, \boldsymbol{\gamma} \rangle < C$ so that the above mathematical program is feasible. It is shown in Dieker et al. (2012) that the solution to (OPT) satisfies $\langle \mathbf{c}, \boldsymbol{\beta} \rangle = C$.

After defining $\tau_i(\boldsymbol{\beta}) := (\beta_i - \gamma_i)\mathbb{E}Z_i^{\boldsymbol{\beta}}$ as noted earlier, the objective function takes the form $t(\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\tau}(\boldsymbol{\beta}))$ for some function $\boldsymbol{\tau}(\boldsymbol{\beta})$, where for $\boldsymbol{\beta} - \boldsymbol{\gamma}, \boldsymbol{w}, \boldsymbol{\tau} > 0$ we have

$$t(\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\tau}) = \sum_{k=1}^{L} w_k \frac{\tau_k}{\beta_k - \gamma_k}. \tag{8}$$

For a queue in a single-class product-form network, $\boldsymbol{\tau}$ is known to be equal to λ and 1 for expected queue length and sojourn time, respectively. Furthermore, τ correspondingly equals $\lambda(c_A^2 + c_S^2)/2$ and $(c_A^2 + c_S^2)/2$ in a single-class Brownian product-form network of GI/GI/1 queues, where c_A^2 and c_S^2 denote the second-order variation terms for the arrival and service process, respectively; see Harrison and Williams (1992). In general stochastic networks, however, $\boldsymbol{\tau}(\boldsymbol{\beta})$ is mathematically intractable.

Our approach relies on the idea that $\boldsymbol{\beta} \mapsto t(\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\tau}(\boldsymbol{\beta}))$ can be reasonably approximated locally by $\boldsymbol{\beta} \mapsto t(\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\tau}(\bar{\boldsymbol{\beta}}))$ in the neighborhood of a given point $\bar{\boldsymbol{\beta}}$. Through this functional form, the i-th term in the approximating objective function only depends on $\boldsymbol{\beta}$ through the one-dimensional quantity β_i, thus effectively "decomposing" the objective function. The explicit incorporation of $\beta_i - \gamma_i$ in the denominator is motivated by the aforementioned product-form results, which effectively result from one-dimensional queueing formulas. We note that the idea of locally approximating the objective function is a well-known principle in trust-region based optimization; refer to, e.g., Conn et al. (2000). Our approach, however, differs significantly from traditional trust-region methods in the motivation, method and analysis.

The following lemma, which is readily proved by applying standard Lagrangian methods, then becomes an essential ingredient in our analysis.

Lemma 2 The minimum of $t(\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\tau})$ over the feasible region in (OPT) is $\boldsymbol{\beta}^*(\boldsymbol{w}, \boldsymbol{\tau})$, where for $\ell = 1, \ldots, L$ $\beta_\ell^*(\boldsymbol{w}, \boldsymbol{\tau}) = \gamma_\ell + (C - \langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle) \frac{\sqrt{w_\ell \tau_\ell / c_\ell}}{\sum_{k=1}^{L} \sqrt{w_k \tau_k c_k}}$.

As an extension of the idea that queue lengths may be approximated locally by functions of the form in (8), and as noted in Section 2, we use the capacity allocation $\boldsymbol{\beta}^*$ determined through the following system of nonlinear equations as the second-phase solution of our approach: For $\ell = 1, \ldots, L$, $\beta_\ell^* = \gamma_\ell + (C - \langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle) \frac{\sqrt{w_\ell \tau_\ell(\boldsymbol{\beta}^*)/c_\ell}}{\sum_{i=1}^{L} \sqrt{w_i \tau_i(\boldsymbol{\beta}^*) c_i}}$. Dieker et al. (2012) show that this system of equations is guaranteed to have a unique solution for a certain precisely defined class of stochastic networks. In an attempt to numerically find a vector $\boldsymbol{\beta}^*$ that satisfies the above equation, assuming existence, one can use the fixed-point iteration scheme with iterates $\{\boldsymbol{\beta}^{(k)} : k \geq 0\}$ given by

$$\beta_\ell^{(k+1)} = \gamma_\ell + (C - \langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle) \frac{\sqrt{w_\ell \tau_\ell(\boldsymbol{\beta}^{(k)})/c_\ell}}{\sum_{i=1}^{L} \sqrt{w_i \tau_i(\boldsymbol{\beta}^{(k)}) c_i}} \quad \text{or} \quad \frac{\beta_i^{(k+1)} - \gamma_i}{\beta_j^{(k+1)} - \gamma_j} = \sqrt{\frac{\beta_i^{(k)} - \gamma_i}{\beta_j^{(k)} - \gamma_j} \times \frac{w_i \mathbb{E}Z_i^{\boldsymbol{\beta}^{(k)}}/c_i}{w_j \mathbb{E}Z_j^{\boldsymbol{\beta}^{(k)}}/c_j}}, \tag{9}$$

The second equation is implied by the first, establishing an important connection with a resource capacity iteration scheme based on observed queue-length information. Since we must allocate at least capacity γ_i to station i, $(\beta_i - \gamma_i)/(\beta_j - \gamma_j)$ is the ratio of "additional" resource capacities allocated to station i and j, respectively. Equation (9) expresses this ratio in terms of the ratio of mean queue lengths, so that more capacity is allocated in the next iterate to stations with disproportionally long queue lengths in the current iterate. The right-hand side of (9) may be interpreted as the geometric mean of two fractions. The geometric average in our case arises from the assumed functional form (8). The effect of building in the asymptote into our algorithm is that the iterates avoid the boundary.

The key approximation is the separable functional form $\tau/(\beta-\lambda)$, which constitutes a nearly universal phenomenon in stochastic networks under a wide range of queueing dynamics. We therefore expect that resource capacity management optimization through an iterative algorithm based on ratios of observed queue lengths and slow-down via geometric means is promising for many different settings. For instance, given a discrete decision space in which to allocate a number of servers to each station, one could use Lemma 2 together with a local search algorithm over the discrete space to generate an iterative scheme. Another interesting variant is the dual formulation of the problem discussed earlier, where the aim is to minimize the total expenditure subject to a bound on the sum of the weighted expected queue lengths.

4 STAGE 2 METHODOLOGY

The second stage of our solution approach uses the latest simulation-based optimization techniques. Here, the literature may be broadly divided into methods that use a broad spectrum of metaheuristics (e.g., tabu search, scatter search, neural networks) to control a sequence of simulation runs in order to find an optimal solution and those that apply several direct methods (e.g., stochastic approximation) which have been widely studied to address simulation-based optimization problems with a more rigorous theoretical foundation.

The metaheuristic approach is often the method of choice for major commercial simulation software products that support optimization.At each step of these metaheuristics, the control procedure selects a new set of candidate optimal solutions by comparing between the simulation results for the current set of decision variables and previously evaluated solution. The choice of the next set of solutions is often motivated by a philosophy of randomized search that is independent of specific problem structure. This structure-agnostic approach gives these methods their greatest appeal (that of being applicable very generally) but is also the source of their greatest weakness, namely the long runtimes that can be incurred in problems of even a modest dimension; see e.g. Heching and Squillante (2013).

The stochastic approximation algorithm for simulation-based optimization has been extensively studied in great generality with rigourous results available on the rates of convergence under reasonable conditions for the objective function. These iterative schemes are effectively the "stochasticization" of a Newton-type iterative optimization (or root-finding) algorithm. Suppose the objective function of a resource allocation optimization problem can be denoted by $z(\boldsymbol{\beta})$. The stochastic approximation iterative algorithm to solve the optimization problem $\min_{\boldsymbol{\beta}} z(\boldsymbol{\beta})$ is $\boldsymbol{\beta}^{(n+1)} = \boldsymbol{\beta}^{(n)} - \varepsilon_n \mathbf{K}\mathbf{Y}^{(n+1)}$, where the variable $\mathbf{Y}^{(n+1)}$ is an estimator of the gradient of $z(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, and the best scaling matrix $\mathbf{K}$ to use is the Hessian of $z(\boldsymbol{\beta})$ at the optimal solution, just as prescribed for Newton-Ralphson type iterative schemes. These methods are regrettably not as common in practice as the metaheuristic approach. One stumbling block has been that the method requires the setting of the critical parameters ε_n to "good" values in order to realize an efficient implementation, where practitioner experience demonstrates that "good" values typically depend on each instance of the problem being solved. Other significant implementation issues surround the efficient and consistent estimation of $\mathbf{Y}$, the gradient of $z(\boldsymbol{\beta})$, as well as its Hessian $\mathbf{K}$. So, though these methods have strong theoretical underpinnings, their use in practice is limited.

5 COMPLETE METHODOLOGY: EXAMPLES

Our solution approach consists of a first-stage analytical-based methodology, comprised of two phases of stochastic approximations, whose results are then used as a starting point for a second-phase simulation-based methodology. We provide two examples of the application of this two stage methodology to solve resource allocation problems in two very different business contexts, both of which demonstrate that our general approach provides optimal solutions at least as good as purely advanced simulation-based optimization methods while taking several orders of magnitude less time to find these optimal solutions.

In the first example, studied in detail in Heching and Squillante (2013), an IT service delivery center is challenged with responding to customer requests that arrive from geographically dispersed customers with widely varying request arrival patterns, service times, and service level agreements. Different agents

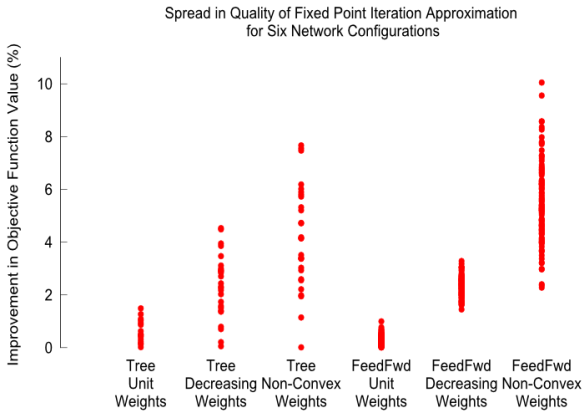

Figure 2: Quality of fixed point iteration result is within 5%, typically 1% for convex settings. Non-Convex cases can have a gap of up to 10%, while average is around 4-5%.

available to respond to requests are capable of serving different subsets of the requests, depending upon their skill. The two-stage methodology is instantiated in this example with the first stage being represented by a first-phase strong-approximation of this optimization problem. In this problem setting, the second phase was not required for the first stage. The metaheuristic Tabu-search method provided the second-stage methodology. Numerical experiments across a wide variety of real-world scenarios demonstrate that leveraging our first-phase results as a starting point provides the same optimal capacity decisions as purely advanced simulation-based optimization methods but with several orders of magnitude reduction in time and resources. In many cases, the first-phase solution for each type of resource differs by at most one from the optimal solution for each work shift and requires four to five orders of magnitude less processing time.

In the second example, studied in detail in Dieker et al. (2012), resource allocation is considered in a stochastic network setting where tasks (or jobs) can be routed between various processing nodes of the network. The underlying stochastic network structure representing the business process for this second example is very different from that of the first example. In this setting, the method treats business processes with a different form of complex network structure. This resource allocation problem is applicable to a variety of stochastic network settings that arise naturally in a variety of data centers that serve traffic from the Internet, as well as some canonical business processes. One of the problem instances considered is based on a tree-like network structure of three tiers of servers that together support large data centers. Our overall methodology is applied to this problem by using the product-form approximation to this stochastic network as the first phase of the first stage, then generalizing to the simulation-evaluation-dependent functional form for the objective function as outlined in Section 3 in the second phase of the first stage, and finally employing stochastic approximation as the method from the second phase of the algorithm.

Recall that the key step in our fixed-point approximation method is the estimation of the average queue lengths of the servers under the capacity values for the current iterate. A simulation-based implementation of queue-length estimation in the fixed-point iteration under the original stochastic network settings yields a consistent estimation. Our results are generated from this simulation-based implementation. Our second-phase method is evaluated based on comparing and contrasting against both the second stage of our solution framework and an approach based solely on the stochastic approximation (SA) algorithm, each of which identifies solutions that are locally optimal. Furthermore, the second phase search for the local optimal solution close to the limit point is obtained by starting the SA algorithm from the limit point identified by the fixed-point iteration algorithm.

The two-phase first-stage procedure was run for each of six network and parameter settings over multiple combinations of coefficients of variations (CoVs) for the interarrival and service distributions. Figure 5 plots the observed "optimality gap" of the limit point identified by the first stage of the two-stage framework

that uses the fixed-point iteration (9) by comparing it against the locally optimal solution identified by the second-stage procedure that uses SA started off from that limit point. It is evident that the SA algorithm is able to improve only by 5% at worst for the cases where the problem is convex, and the additional improvement is in most cases only about 1.0-1.5%. The performance of the algorithm degrades a bit for the non-convex case, with the worst case improvement rising to about 10% and the average case performance being in the 3-5% range. In contrast, the objective function value of the optimal capacity allocation obtained from the Jackson product-form approximation, that is, setting all CoVs to 1, was observed to have a relative optimality gap of between 75% and 350%, clearly indicating the poor quality of this simplistic assumption.

Convex optimization problem settings have unique globally optimal solutions, and thus the fixed point identified in the first phase of our approach is unique when (OPT) is convex. Under the parameter values considered, whenever the server weights satisfy $w_{\pi(i)} \geq w_i$ the Brownian tree network version of the optimization problem (OPT) is known to be convex (Dieker, Ghosh, and Squillante 2012). For the same network configuration when the weights w_i yield a difference-of-convex functions for the objective function of (OPT), the problem may have multiple local optima and our fixed-point iteration algorithm itself may have multiple limit points.

The second phase of the first stage method proposed in this paper finds approximations to the local optimal solutions that are on average within 5% of optimality gap. In addition, the method is parameter-less and insensitive to the stopping criterion chosen for the simulations. This provides major savings in computational time as compared to the second stage method being employed for the whole search. Regardless, all stochastic approximation algorithms take one or more orders of magnitude to converge compared to the method proposed in this article. In addition, the key savings are also relatable to the lack of any parameters in the fixed-point scheme that the user must tweak for faster convergence.

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. Springer.

Boyd, S., and L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Chen, H., and D. D. Yao. 2001. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer-Verlag.

Conn, A. R., N. I. M. Gould, and P. L. Toint. 2000. *Trust-region methods*. Philadelphia,PA: SIAM.

Dai, J. G., and J. M. Harrison. 1992. "Reflected Brownian motion in an orthant: numerical methods for steady state analysis". *Advances in Applied Probability* 2:6–86.

Dieker, A. B., S. Ghosh, and M. S. Squillante. 2008. "Capacity Optimization in Feedforward Brownian Networks". *Performance Evaluation Review* 36 (2).

Dieker, A. B., S. Ghosh, and M. S. Squillante. 2012. "Optimal Resource Capacity Management for Stochastic Networks". Preprint.

Harrison, J. M., and R. J. Williams. 1987. "Brownian models of open queueing networks with homogeneous customer populations". *Stochastics* 22:77–115.

Harrison, J. M., and R. J. Williams. 1992. "Brownian Models of Feedforward Queueing Networks: Quasireversibility and Product Form Solutions". *Annals of Applied Probability* 2 (2): 263–293.

Heching, A. R., and M. S. Squillante. 2013. "Optimal Capacity Management and Planning in Services Delivery Centers". Preprint.

Menasce, D. A., and V. A. Almeida. 2000. *Scaling for E-Business: Technologies, Models, Performance, and Capacity Planning*. Prentice Hall.

Nelson, B. L., and S. G. Henderson. 2007. *Handbooks in OR and MS: Simulation*. Elsevier Science.

Squillante, M. S. 2011. "Stochastic Analysis and Optimization of Multiserver Systems". In *Run-time Models for Self-managing Systems and Applications*, edited by D. Ardagna and L. Zhang, Chapter 1, 1–24. Springer.

AUTHOR BIOGRAPHIES

SOUMYADIP GHOSH is a Research Staff Member in the Business Analytics and Mathematical Sciences Department at the IBM T.J. Watson Research Center. His current research interests lie in simulation based optimization techniques for stochastic optimization problems, with a focus on applications in Energy and Power systems and supply chain management. His email is ghoshs@us.ibm.com and his web page is at https://researcher.ibm.com/researcher/view.php?person=us-ghoshs.

ALIZA HECHING is a Research Staff Member in the Mathematical Sciences Department at the IBM T.J. Watson Research Center. Her research interests include modeling, analysis, and optimization with a current focus on optimal workforce management and the analysis and design of service systems. Her email address is ahechi@us.ibm.com and her web page is http://researcher.watson.ibm.com/researcher/view.php?person=us-ahechi.

MARK S. SQUILLANTE is a Research Staff Member and Manager in the Mathematical Sciences Department at the IBM T.J. Watson Research Center. His research interests broadly concern mathematical foundations of the analysis, modeling and optimization of complex stochastic systems. He is a Fellow of ACM and IEEE, a member of AMS, Bernoulli Society, IMS, INFORMS and SIAM, and serves on the Editorial Boards of *Operations Research*, *Performance Evaluation* and *Stochastic Models*. His email address is mss@us.ibm.com and his web page is http://researcher.ibm.com/researcher/view.php?person=us-mss.