RC24241 (W0704-078) April 13, 2007
Mathematics

IBM Research Report

Optimal Capacity Planning in Stochastic Loss Networks with Time-Varying Workloads

Sandeep Bhadra
Department of Electrical and Computer Engineering
University of Texas
Austin, TX 78712

Yingdong Lu, Mark S. Squillante
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

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

Optimal Capacity Planning in Stochastic Loss Networks with Time-Varying Workloads

Sandeep Bhadra
Dept. of Electrical and Computer Engineering
University of Texas
Austin, TX 78712, USA
sandeepb@mail.utexas.edu

Yingdong Lu and Mark S. Squillante
Mathematical Sciences Department
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA
{yingdong,mss}@watson.ibm.com

ABSTRACT

We consider a capacity planning optimization problem in a general theoretical framework that extends the classical Erlang loss model and related stochastic loss networks to support time-varying workloads. The time horizon consists of a sequence of coarse time intervals, each of which involves a stochastic loss network under a fixed multi-class workload that can change in a general manner from one interval to the next. The optimization problem consists of determining the capacities for each time interval that maximize a utility function over the entire time horizon, finite or infinite, where rewards gained from servicing customers are offset by penalties associated with deploying capacities in an interval and with changing capacities among intervals. We derive a state-dependent optimal policy within the context of a particular limiting regime of the optimization problem, and we prove this solution to be asymptotically optimal. Then, under fairly mild conditions, we prove that a similar structural property holds for the optimal solution of the original stochastic optimization problem, and we show how the optimal capacities comprising this solution can be efficiently computed.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Stochastic processes, Markov processes, Queueing theory; G.1.6 [**Optimization**]: Linear programming, Convex programming

General Terms

Theory, Algorithms, Performance

Keywords

Stochastic loss networks, Stochastic dynamic programming, Asymptotic optimality, Erlang loss formula, Erlang fixed-point approximation, Capacity planning, Time-varying workloads

1. INTRODUCTION

Many resource allocation problems, which generally involve resources of different capabilities that are used to provide service to various classes of users for finite periods of time (often randomly distributed due to inherent variability and uncertainty), can be modeled and analyzed as network problems. The connection stems from the observation that in a network with fixed source-destination pairs (users) and fixed routes (combinations of resources required for service), satisfying the traffic requirements of each user amounts to utilizing a certain fraction of the capacity of each of the resources (links) on the source-destination route. In particular, stochastic loss networks have been well established as an effective framework for modeling and analyzing the allocation of multiple resources among multiple classes of non-backlogging workloads for almost a century. The classical Erlang formula, which provides the probabilistic characterization of a stochastic loss network, has been thoroughly studied and applied in a highly diversified set of research fields. Due to the computational complexity of the exact Erlang formula and related measures, a fixed-point approximation has been proposed and extensively studied as an efficient numerical procedure for calculating performance metrics associated with the stochastic loss network, which in turn has been successfully used in various applications of stochastic loss networks.

On the other hand, there are a large number of applications in various fields that require the solution of resource allocation problems over relatively long time horizons during which the multi-class workload varies among relatively coarse intervals comprising the time horizon. This is generally referred to as the class of resource capacity planning problems and they arise in a wide variety of application domains such as telephony and data networks, distributed computing and data centers, inventory control and manufacturing systems, and call and contact centers, just to name a few. The traditional Erlang loss model and its fixed-point approximations do not support the type of time-varying workloads encountered in these important application areas. Furthermore, some fundamental issues of theoretical and practical importance are encountered upon extending stochastic loss networks and the domains to which they can be applied. This includes characterizing how performance and optimization results vary under forms of non-stationary workload processes, understanding the degree to which the Erlang formula and some of its variants can be and need to be effectively extended, and characterizing the relationship between the loss probabilities and other performance metrics.

To address these fundamental issues and to model and solve a general class of resource capacity planning optimization problems, we consider two key extensions of classical stochastic loss networks. First, we allow the multi-class workload to vary over time. More specifically, we consider the resource capacity planning optimization problem in a more general theoretical framework that consists of a sequence of relatively coarse time intervals, or epochs, each of which is comprised of a stochastic loss network under a

.

fixed multi-class workload, and where this multi-class workload can change in a general manner from one epoch to another. This is achieved by introducing a corresponding sequence of Markov-modulated processes that govern the state of the multi-class, multi-epoch workload process and the changes in this state of the workload process at the end of each epoch. All epochs, possibly of different lengths, are relatively coarse and sufficiently long for the stochastic process modeling the loss network and its performance metrics to reach stationarity within each epoch. The multiple time scales involved in the applications of interest provide both theoretical and practical support for our stationary stochastic approach. Hence, our general framework makes it possible for the stationary performance results of stochastic loss networks, including the Erlang formula and its fixed-point approximations, to continue to be applied in order to characterize the behavior of the stochastic loss network within each epoch comprising the time horizon.

Second, in addition to the performance analysis aspects of our general framework as extensions of previous work in the literature, we seek to determine the optimal capacity levels for each epoch with respect to an objective function defined over the entire time horizon. More specifically, the resource capacity planning optimization problem consists of determining the resource capacities throughout each epoch that maximize the expectation of a utility function over the time horizon of interest, which may be finite or infinite. A global utility function is based, without loss of generality, on expected utility in which rewards are gained for servicing customers who are accepted into the system at the time of their arrival and penalties are incurred as the result of deploying resource capacity levels throughout an epoch and as the result of increasing or decreasing capacity levels from one epoch to the next. We therefore formulate and solve a stochastic dynamic programming problem to determine the optimal capacities throughout each epoch that maximize the average utility over time.

Our optimization results are based on two different forms of analysis, one based on asymptotic properties and the other based on general structural properties. By exploiting and extending the limiting regime theory developed by various researchers, most notably Kelly [15, 16] and Whitt [23] (also refer to [17]), we derive a state-dependent threshold-based capacity planning policy within the context of the limiting regime through linear programming and we prove that this policy is asymptotically optimal. Then, under quite mild conditions on the general form of the utility function which often hold in the application domains motivating our study, we prove that the optimal solution of the corresponding stochastic dynamic program has similar threshold-type structural properties for the capacity levels, and we derive a procedure based on these results for efficiently calculating the optimal capacity levels. Our results are established for both the finite-horizon and infinite-horizon instances of the stochastic dynamic program.

While the problems we consider are important from the theoretical perspectives of investigating stochastic loss networks in general, and Erlang fixed-point approximations in particular, and of investigating the corresponding optimal resource allocation and stochastic control problems, our analysis and results can support a wide range of practical resource allocation problems in various application domains involving time-varying workloads. In such cases, we can model the network problems in terms of a sequence of time intervals at relatively coarse time scales during which the workload process is stationary and across which the workload process can change in a general manner governed by a sequence of Markov-modulated processes. For example, consider the case of a wireless network operator that is required to satisfy calls of various types (viz. data, voice, and video) over a common set of links with finite bandwidth constraints. As the mean traffic intensities of calls in data, voice and video vary with time in each day and from weekdays to weekends, the natural concern of the operator would be to appropriately allocate these bandwidth resources over the various call-classes so as to maximize net income. The related problem of resource reservation and allocation for data networks has been studied in a different context for TCP call book-ahead by Greenberg, Srikant and Whitt [13] and Wischik and Greenberg [25].

Another emerging application area motivating our study is workforce management where, e.g., an information technology (IT) services company offers a collection of IT service products (routes) each of which requires a set of resources with certain capabilities (links). The customer demand for such IT service products can vary over relatively coarse intervals of time and the IT services company seeks to maximize its profits over a long-term horizon comprised of multiple instances of these time intervals. Assuming the product offerings remain fixed over the time horizon of interest (which can be relaxed as part of future work), the IT services company can adjust its per-class resource capacities in response to the time-varying multi-product workloads in order to maximize its profits over the long run. A related problem in multi-item inventory systems has also received attention in the research literature [26, 19].

Related Work. Historically, the Erlang formula [11] has been used to study the drop rate of calls in busy telephone networks with constrained link capacities. Given a telephone network and a prediction of expected demand, the formula estimates the steady-state probability that a call will be dropped by the network. While the initial results of Erlang were for the particular case of Poisson arrivals and exponential durations, the results of Sevastyanov [22] demonstrate that the Erlang formula holds even in the presence of non-exponential finite-mean distributions for the call durations. The results are known to hold even in the presence of dependencies among duration times for a particular route (cf. Burman et al. [9]). The recent results of Bonald [8] further suggest that relaxations can be made to the call arrival process, merely requiring that users generate sessions according to a Poisson process and, within each session, blocked users may retry with a fixed probability after an idle period of random length. Under these mild conditions, the Erlang formula is known to be insensitive to all traffic characteristics beyond the traffic intensity such as the number of calls per session, the call duration times or the idle time distribution. These results suggest that the Erlang formula is an effective and sufficiently accurate way to model the loss behavior of networks with finite resources.

On the other hand, the exact calculation of the loss probability, for example, the Erlang formula, is of limited use in large networks since its computation is known to be $\sharp P$-complete in the size of the network [18] and thus computationally intractable for many of the applications motivating our study. An Erlang fixed-point approximation of the loss probabilities has been developed to address this computational complexity through a product-form expression of the blocking probabilities on the individual links comprising a route [23, 15, 16]. In other words, it is as if the call loss is caused by independent blocking events in each of the links on the route. Kelly [15, 16] considers the set of implicit functions mapping the blocking probability of each link to the blocking probabilities of other links via the Erlang function. The Erlang fixed-point approximation provides a tractable way of estimating the loss probabilities in large networks. Furthermore, a few researchers, including Kelly [15, 16] and Whitt [23] (also refer to [17]), have proven that in the limiting regime as the traffic intensities and link capacities grow together in a proportional manner, the Erlang fixed-point approximation is asymptotically exact.

Summary of Contributions. We extend traditional stochastic loss networks, including the Erlang model and its fixed-point approximations, to support multi-class workloads that vary over relatively coarse intervals of time. A general systematic framework is developed for capacity planning optimization problems that includes: Asymptotic properties of optimal resource allocation solutions and proving the asymptotic optimality of these stochastic network control solutions in the limiting regime; Structural properties of optimal stochastic dynamic programming solutions and proving the general optimality of these stochastic network control solutions under mild conditions on the objective function; General optimality results that go beyond stochastic loss networks and even beyond resource capacity planning problems. A collection of numerical results have been obtained that confirm and quantify our theoretical results.

Paper Organization. Section 2 presents our general theoretical framework and some preliminary technical results. The main results of the paper are then provided in Sections 3 and 4, which consider our asymptotic and general resource allocation problems, respectively. Section 5 presents a representative sample of our numerical results, and Section 6 provides concluding remarks.

2. TECHNICAL PRELIMINARIES

In this section we present notational conventions used throughout the paper, the stochastic loss models and stochastic optimization problems considered in our study, and some preliminary technical results. Let $\mathbb{Z}^+$ and $\mathbb{Z}_+$ denote the set of positive and nonnegative integers, respectively, with $\mathbb{R}^+$ and $\mathbb{R}_+$ denoting the corresponding sets of reals. Bold letters shall be used for matrices and vectors. Column vectors are assumed unless noted otherwise. The transpose of a matrix or vector $\mathbf{M}$ shall be denoted by $\mathbf{M}^T$.

We consider a stochastic loss network consisting of a set of links, indexed by $j = 1, \ldots, J$, and a set of routes $\mathcal{R}$ that are defined as a collection of links and are indexed by $r = 1, \ldots, |\mathcal{R}|$. A call for route r requires capacity A_{jr} from link j, $A_{jr} \in \mathbb{Z}_+$, where each link j has capacity C_j. We assume that calls for route r arrive from an independent Poisson process with rate λ_r. Such a call arrival is blocked and lost if the available capacity on any link j is less than A_{jr}, $\forall j = 1, \ldots, J$, and otherwise the call reserves the available capacity A_{jr} on each link j for a duration following a general distribution with mean μ_r^{-1}, $\forall j = 1, \ldots, J$. The traffic intensity for route r is denoted by $\nu_r = \lambda_r / \mu_r$. Call arrival and duration times are assumed to be mutually independent. It is important to note that many of our results can be extended to handle relaxations of some of the above assumptions of the Erlang loss model, e.g., see [8], and in particular the results of Section 4 hold for general stochastic networks (not even limited to loss networks) under the assumptions specified therein.

Define

$$\begin{aligned} \mathbf{A} &\triangleq [A_{jr}]_{j=1,\ldots,J; r=1,\ldots,|\mathcal{R}|}, \\ \mathbf{C} &\triangleq (C_1, C_2, \ldots, C_J), \\ \boldsymbol{\nu} &\triangleq (\nu_1, \ldots, \nu_{|\mathcal{R}|}). \end{aligned}$$

Let M_r denote the number of active calls using route r and define $\mathbf{M} \triangleq (M_1, \ldots, M_{|\mathcal{R}|})$. Then it is well known that $\mathbf{M}$ has a unique stationary distribution π and it is given by

$$\pi(\mathbf{m}) \;=\; G(\mathbf{C}) \prod_{r=1}^{|\mathcal{R}|} \frac{\nu_r^{m_r}}{m_r!}, \qquad \mathbf{m} \in \mathcal{S}(\mathbf{C}), \tag{1}$$

where

$$\mathcal{S}(\mathbf{C}) \;=\; \{\mathbf{n} \in \mathbb{Z}_+^{|\mathcal{R}|} : \mathbf{Am} \le \mathbf{C}\} \tag{2}$$

and $G(\mathbf{C})$ is the normalization factor

$$G(\mathbf{C}) \;=\; \left(\sum_{\mathbf{m} \in \mathcal{S}(\mathbf{C})} \prod_{r=1}^{|\mathcal{R}|} \frac{\nu_r^{m_r}}{m_r!} \right)^{-1}. \tag{3}$$

Further, the stationary probability that a call for route r is lost can be expressed as

$$L_r \;=\; 1 \;-\; G(\mathbf{C})^{-1} G(\mathbf{C} - \mathbf{A}\mathbf{e}_r), \tag{4}$$

where $\mathbf{e}_r \in \mathcal{S}(\mathbf{C})$ is the unit vector corresponding to a single active call using route r. Refer to, e.g., [17] and the references therein.

One can obviously see from (1) – (3) that the computational complexity of calculating the exact value of $G(\mathbf{C})$, and in turn the exact stationary distribution, even for moderate values of J and $|\mathcal{R}|$, grows very quickly and this causes such calculations to be computationally intractable. In fact, calculating $\pi(\mathbf{m})$ is known to be $\sharp P$ complete [18]. This computational complexity has been a primary motivation for the well-known Erlang fixed-point approximation in which the stationary loss probabilities L_r for routes r are given by

$$L_r \;=\; 1 \;-\; \prod_{j=1}^{J} (1 - B_j)^{A_{jr}}, \tag{5}$$

where the blocking probabilities B_j for links j satisfy the system of nonlinear equations

$$B_j \;=\; E\left((1 - B_j)^{-1} \sum_{r=1}^{|\mathcal{R}|} A_{jr} \nu_r \prod_{i=1}^{J} (1 - B_i)^{A_{ir}},\; C_j \right), \tag{6}$$

with

$$E(\nu, C) \;=\; \frac{\nu^C}{C!} \left(\sum_{n=0}^{C} \frac{\nu^n}{n!} \right)^{-1} \tag{7}$$

being the Erlang formula for the loss probability of an isolated link of capacity C under traffic from a Poisson stream of intensity ν. The corresponding effective traffic intensity for route r is given by

$$\gamma_r \;=\; (1 - L_r) \nu_r.$$

Define

$$\begin{aligned} \mathbf{L} &\triangleq (L_1, \ldots, L_{|\mathcal{R}|}), \\ \mathbf{B} &\triangleq (B_1, \ldots, B_J), \\ \boldsymbol{\gamma} &\triangleq (\gamma_1, \ldots, \gamma_{|\mathcal{R}|}). \end{aligned}$$

It is well-known that there exists a solution $\mathbf{B} \in [0, 1]^J$ of the Erlang fixed-point equations (6) and that this solution converges to the exact solution (1) – (3) of the Erlang loss model in the limit as the traffic intensity vector $\boldsymbol{\nu}$ and capacity vector $\mathbf{C}$ are increased together in fixed proportion; see [23, 15]. The asymptotic exactness of the Erlang fixed-point approximation follows from an instance of the central limit theorem for conditional Poisson random variables in which $\boldsymbol{\nu}$ and $\mathbf{C}$ grow together. Namely, the $|\mathcal{R}|$ Poisson random variables being truncated by a polytope involving the capacities $\mathbf{C}$ are approximated by $|\mathcal{R}|$ independent normal random variables truncated by the polytope.

Motivated by the applications and discussion in the introduction, we now extend the foregoing stochastic loss network framework (including its notation based on a single stationary interval) to support time-varying multi-class workloads. Our general theoretical framework is based on a sequence of time intervals, or epochs, where: (i) each epoch consists of a stochastic loss network with fixed workload and capacity vectors, including those presented above; (ii) the workload can change from one epoch to the next; and (iii) the capacity vectors are the control variables of interest that can be adjusted in response to such workload changes. We consider a time horizon consisting of $N \geq 1$ epochs, indexed by $n = 0, \ldots, N-1$, with N finite or infinite. The dynamics of the time-varying workload is governed by a sequence of Markov-modulated processes with transition probability matrices $\mathbf{P}_n$ of order $U < \infty$ and an initial probability vector $\boldsymbol{\alpha}$. (For convenience, we shall refer to the Markov-modulated process $\mathbf{P}_n$.) More specifically, the system starts epoch $n = 0$ in state i with probability $\boldsymbol{\alpha}(i)$, remains in state i for a length of time T_0, and then at the end of epoch $n = 0$ the system transitions to state i' with probability $\mathbf{P}_{0,ii'}$, for all $i, i' = 1, \ldots, U$. These dynamics continue for epoch $n = 1, \ldots, N-1$ starting in any state i where the system remains in state i for a length of time T_n and at the end of epoch n the system transitions to state i' with probability $\mathbf{P}_{n,ii'}$, for all $i, i' = 1, \ldots, U$.

The length T_n of each epoch n is assumed to be sufficiently large in comparison with the mean interarrival and duration times of calls. Under this assumption, it is known that the time for the system to reach stationarity is also small in comparison with T_n, and hence stationary statistics can be used to closely approximate the system behavior at any time within the interval. See, e.g., [12] and [24] for a detailed analysis of the accuracy of these assumptions in similar stochastic systems. Furthermore, from a different perspective, the modulating Markov process and the call arrival and duration times can be treated as processes of different time scales, for which the multi-scale limit theorems in [2] and [27] have demonstrated the efficiency and accuracy of stationary approaches such as ours across a wide range of stochastic systems. From a practical perspective, the various applications motivating our study involve changes in multi-class workloads at time scales much coarser than the interarrival and duration times of customer requests, where resource capacity adjustments in response to workload changes can be made at relatively finer time scales. Meanwhile, resource capacity allocations tend to be stable if the workload does not change.

Throughout each epoch n in any state $i = 1, \ldots, U$ the stochastic loss network has (nonnegative) input parameters

$$\begin{aligned} \mathbf{A}(i) &\triangleq [A_{jr}(i)]_{j=1,\ldots,J;r=1,\ldots,|\mathcal{R}|}, \\ \mathbf{C}_n(i) &\triangleq (C_{n,1}(i), \ldots, C_{n,J}(i)), \\ \boldsymbol{\nu}_n(i) &\triangleq (\nu_{n,1}(i), \ldots, \nu_{n,|\mathcal{R}|}(i)) \end{aligned}$$

under which the stochastic network yields the (nonnegative) results

$$\begin{aligned} \mathbf{L}_n(i) &\triangleq (L_{n,1}(i), \ldots, L_{n,|\mathcal{R}|}(i)), \\ \boldsymbol{\gamma}_n(i) &\triangleq (\gamma_{n,1}(i), \ldots, \gamma_{n,|\mathcal{R}|}(i)), \\ \mathbf{B}_n(i) &\triangleq (B_{n,1}(i), \ldots, B_{n,J}(i)). \end{aligned}$$

We note that the sequence of stochastic transition probability matrices $\mathbf{P}_n$, when $N < \infty$, can be stationary or nonstationary, periodic or aperiodic, recurrent or transient, and so on; our sole assumption in this case is that each $\mathbf{P}_n$ is stochastic. On the other hand, when N is infinite, we assume that $\mathbf{P}_n = \mathbf{P}$ for all n, that $\mathbf{P}$ is ergodic, and that $\boldsymbol{\nu}_n(i) = \boldsymbol{\nu}(i)$ for all n and all $i = 1, \ldots, U$.

Our general theoretical framework is also based on a sequence of stochastic optimization problems to determine the capacity vectors that should be deployed within each epoch n and across epochs in order to maximize a utility function over the entire time horizon, where rewards (revenues) are gained for accepted calls, penalties (costs) are incurred for the capacity vector deployed, and penalties (costs) are incurred for changes in the capacity vector. We formulate this net-utility maximization problem as a stochastic dynamic program over the horizon of N epochs in terms of the following revenue and cost functions. Let $R_n(i, \mathbf{C}_n)$ denote the expected revenue rate for epoch n in state i given a capacity requirement matrix $\mathbf{A}(i)$, a traffic intensity vector $\boldsymbol{\nu}_n(i)$ and a capacity vector $\mathbf{C}_n$. Further let $\mathbf{K}$ denote the J-vector of cost rates for link capacities, $\mathbf{I}$ the J-vector of cost rates for increasing link capacities, and $\mathbf{D}$ the J-vector of cost rates for decreasing link capacities.

In our main formulation of interest, we assume that the revenue is linear with respect to the number of calls accepted, that the expected revenues for epoch n in each state i are given by

$$R_n(i, \mathbf{C}_n) T_n \;=\; \mathbf{w} \cdot \boldsymbol{\gamma} T_n, \tag{8}$$

and that the expected profits for epoch n in state i are expressed as

$$P_n(i) \;=\; R_n(i, \mathbf{C}_n) T_n - (\mathbf{K} \cdot \mathbf{C}_n T_n + \mathbf{I} \cdot \mathbf{x}_n^+ + \mathbf{D} \cdot \mathbf{x}_n^-), \tag{9}$$

where

$$\mathbf{C}_{n+1} \;=\; \mathbf{C}_n + \mathbf{x}_n, \tag{10}$$

$\mathbf{w}$ is the $|\mathcal{R}|$-vector of base revenue rates, $\mathbf{x}^+ = (x_1^+, \ldots, x_J^+)$ with $x^+ = \max\{x, 0\}$, and $\mathbf{x}^- = (x_1^-, \ldots, x_J^-)$ with $x^- = \max\{-x, 0\}$. For each epoch, the expected revenues as a function of the capacities deployed throughout the epoch depend upon the structural properties of $\boldsymbol{\gamma}$, whereas the expected costs consist of a linear function of the capacities deployed throughout the epoch, a linear function of any increases in capacities, and a linear function of any decreases in capacities. In Section 4, a few of these details of our formulation are relaxed and generalized in order to study an even broader class of capacity planning optimization problems.

Our objective is to maximize the expected profit over the entire time horizon with respect to the capacity vector decision variables that can be adjusted in response to workload changes. The corresponding capacity planning optimization problem based on the expected total discounted profit over the entire time horizon, with discount factor β, is in general given by the formulation

$$\max_{(\mathbf{x}_0, \ldots, \mathbf{x}_{N-1})} \quad \mathbb{E}\left[\sum_{n=0}^{N-1} e^{-\beta n} P_n(i_n)\right] \tag{11}$$

$$\text{s.t.} \quad \mathbf{x} \triangleq (\mathbf{x}_0, \ldots, \mathbf{x}_{N-1}) \geq \mathbf{0}, \tag{12}$$

where the expectation is over $(\Omega, \mathcal{F}, \mathbb{P})$. This stochastic dynamic program is quite general [21, 6] and it can be used to solve the capacity planning optimization problems from a broad spectrum of different applications that fit our general theoretical framework.

The capacity planning policies of interest specify the capacity vectors that should be deployed throughout the entire time horizon including the changes in these capacity vectors over time. Our main results in Sections 3 and 4 will establish that the optimal capacity planning policy solving the stochastic dynamic program (11),(12) has an important structural property, often refered to as a threshold-based policy and in the present context is defined as follows.

DEFINITION 2.1. *A* threshold-based capacity planning policy, *denoted by $\mathbb{C}$, specifies the $N \times J \cdot U$ capacity matrix $\mathbf{C}$ in terms of the $J \cdot U$ capacity vector $\mathbf{C}_0$ that is employed throughout epoch* 0*, the actions to increase and decrease capacities from $\mathbf{C}_0$ to $\mathbf{C}_1$, the $J \cdot U$ capacity vector $\mathbf{C}_1$ that is employed throughout epoch* 1*, the actions to increase and decrease capacities from $\mathbf{C}_1$ to $\mathbf{C}_2$, and so on for all epochs $n = 0, \ldots, N-1$ and all states $i = 1, \ldots, U$. The capacity vector levels are called the* policy thresholds.

Correspondingly, we denote the expectation under any capacity planning policy $\mathbb{C}$ by the operator $\mathbb{E}_{\mathbb{C}}[\cdot]$.

The class of optimal threshold-based policies considered in this paper are related to the classical optimal base-stock policies from inventory management systems, which consist of ordering up to the optimal base-stock level whenever the inventory drops below this level [28]. Our optimal threshold-based policies, however, have a number of important differences with the traditional optimal base-stock policies. This includes the additional flexibility of reducing the capacity vector levels down to the optimal policy thresholds whenever the capacity vector exceeds these thresholds.

3. ASYMPTOTIC OPTIMALITY

In this section we consider the stochastic optimization problem of Section 2 in a particular limiting regime of our original formulation. This limiting regime, proposed by Kelly [15, 16, 17], characterizes the asymptotic dynamics of the stochastic loss network, as well as the special structural properties of the Erlang functions, all within a very efficient framework. This limiting regime has been the standard model for the asymptotic behavior of stochastic loss networks, and the basic ideas have been adapted in the study of general models for stochastic loss networks; see, e.g., [8]. We derive an optimal solution of the stochastic dynamic program in this limiting regime, for both finite and infinite horizons, we establish that the corresponding optimal policies have threshold-based structural properties, and then we prove that this solution is asymptotically optimal with respect to the original formulation of Section 2.

3.1 Stochastic Dynamic Program

Consider the stochastic optimization problem (11),(12) in the limiting regime as the traffic intensity vector $\boldsymbol{\nu}_n$ and capacity vector $\mathbf{C}_n$ for every epoch n are increased together in fixed proportion. Kelly [15, 17] has shown that the blocking probabilities of the stationary single-epoch Erlang loss model in this limiting regime are functions of the solutions of the so-called convex dual problem formulation

$$\begin{array}{ll} \min\limits_{\mathbf{y}} & \displaystyle\sum_{r=1}^{|\mathcal{R}|} \nu_r \exp(-\mathbf{y}^T \mathbf{A} \cdot \mathbf{e}_r) + \mathbf{y} \cdot \mathbf{C} \\ \text{s.t.} & \mathbf{y} \geq \mathbf{0}, \end{array}$$

where $B_j = 1 - \exp(-y_j^*)$ and y^* is the optimum of the convex problem. In fact, this is the dual to the following (primal) convex program

$$\begin{array}{ll} \max\limits_{\mathbf{x}} & \displaystyle\sum_{r=1}^{|\mathcal{R}|} x_r \log \nu_r - x_r \log x_r + x_r \\ \text{s.t.} & \mathbf{x} \geq \mathbf{0}, \quad \mathbf{A}\mathbf{x} \leq \mathbf{C}, \end{array}$$

which is formulated to identify the most probable state in the Erlang loss model. There is a unique optimum x^* of the primal problem that can be expressed in the form

$$x_r^* \;=\; \nu_r \prod_{j=1}^{J} (1 - B_j)^{A_{jr}}$$

where $(B_1, B_2, \ldots, B_J) \in [0,1]^J$ is any solution to

$$\sum_{r=1}^{|\mathcal{R}|} A_{jr} \nu_r \prod_{i=1}^{J} (1 - B_i)^{A_{ir}} \begin{cases} = C_j & \text{if } B_j > 0, \\ \leq C_j & \text{if } B_j = 0, \end{cases} \tag{13}$$

or, equivalently in terms of L_r,

$$\sum_{r=1}^{|\mathcal{R}|} A_{jr} \nu_r (1 - L_r) \begin{cases} = C_j & \text{if } B_j > 0, \\ \leq C_j & \text{if } B_j = 0. \end{cases} \tag{14}$$

The conditions in (14) have the useful interpretation that the capacity C_j of any link j for which $B_j > 0$ must be completely utilized with respect to the superposition of the traffic intensities ν_r thinned by their stationary loss probabilities L_r over all routes $r \in \mathcal{R}$.

Note that equation (14) provides the feasible polytope for $\nu_r(1 - L_r)$ in the limiting regime, and thus, in this form, the dual problem can be treated as a linear program. Then from linear programming theory (e.g., refer to [10, 7]), we know that $\nu_r(1 - L_r)$ is linear in $\mathbf{C}$ which together with (8) establishes the following result.

LEMMA 3.1. *For any epoch $n = 0, 1, \ldots, N-1$ and state $i = 1, 2, \ldots U$ in the limiting regime, the revenue function $R_n(i; \mathbf{C})$ is linear in $\mathbf{C}$.*

Now, for each epoch n, we can write the revenue function as $R_n(i, \mathbf{C}) = \mathbf{Q}_n(i)^T \cdot \mathbf{C}_n$ for some appropriate $\mathbf{Q}_n(i) \in \mathbb{R}^J$, which together with (9) provides the expected profit for epoch n in state i. The associated Bellman optimality equations (see, e.g., [21, 6]) are then given by

$$J_n(i, \mathbf{C}_n) \;=\; \max_{\mathbf{x}_n} P_n(i, \mathbf{C}_n, \mathbf{x}_n) \tag{15}$$

$$\begin{aligned} P_n(i, \mathbf{C}_n, \mathbf{x}_n) \;=\; & e^{-\beta} \sum_{j=1}^{U} \mathbf{P}_{n,ij} J_{n+1}(j, \mathbf{C}_n + \mathbf{x}_n) \\ & + H_n(i, \mathbf{C}_n, \mathbf{x}_n) \end{aligned} \tag{16}$$

for all $n = 0, \ldots, N-1$, where

$$H_n(i, \mathbf{C}, \mathbf{x}) \triangleq R_n(i, \mathbf{C}) T_n - \mathbf{K} \cdot \mathbf{C} T_n - \mathbf{I} \cdot \mathbf{x}^+(i) - \mathbf{D} \cdot \mathbf{x}^-(i)$$

and $J_n(i, \mathbf{C}_n)$ is the value function for maximizing expected profit over the time horizon from epoch n to epoch $N-1$ starting in state i with capacity $\mathbf{C}_n$ assuming $J_N(i, \mathbf{C}_N) = 0$. It is easy to see that $H_n(i, \mathbf{C}, \mathbf{x})$ is a concave function in $\mathbf{C}$ for each $n = 0, \ldots, N-1$. Following a standard recursive argument (refer to, e.g., [21, 6]), it can be shown that $J_n(i, \mathbf{C}_n)$ is also concave in $\mathbf{C}_n$ and that the optimal capacity planning policy is a threshold-based policy.

Let us next turn to determine the optimal capacity vector thresholds, first considering the case where N is finite. The above problem then can be solved using standard dynamic programming algorithms. More importantly, however, based upon the observation that the capacity decisions do not affect the traffic intensity, we can further simplify the problem. Letting $\mathbf{u}_n(i)$ and $\mathbf{d}_n(i)$ denote the positive and negative part of $\mathbf{x}_n(i)$, then the solution to the following linear program will provide the optimal solution to (11),(12).

$$\begin{array}{ll} \max & \displaystyle\sum_{n=0}^{N-1} \sum_{i=1}^{U} e^{-\beta n} \left(\prod_{\ell=0}^{n-1} \mathbf{P}_\ell \right) H_n(i, \mathbf{C}(i), \mathbf{u}(i) - \mathbf{d}(i)) \\ \text{s.t.} & \displaystyle\sum_{r=1}^{|\mathcal{R}|} A_{jr}(i) \nu_{n,r}(i) (1 - L_{n,r}(i)) \leq C_{n,j}(i) \\ & \mathbf{u}(i) \geq 0, \mathbf{d}(i) \geq 0. \end{array}$$

To see this is true, we only need to verify that when $B_{n,j}(i) > 0$ for some n, j, i, the constraint inequalities are tight. Suppose otherwise the existence of a triplet (n, j, i) such that $B_{n,j}(i) > 0$ while

$$A_{jr}(i) \nu_{n,r}(i) (1 - L_{n,r}(i)) < C_{n,j}(i).$$

Define a new capacity vector $\tilde{\mathbf{C}}$ that coincides with $\mathbf{C}$ at each component except (n, j, i), where

$$\tilde{C}_{n,j}(i) = A_{jr}(i) \nu_{n,r}(i) (1 - L_{n,r}(i)).$$

Then we can see that $\tilde{\mathbf{C}}$ is a feasible solution, but it yields a higher objective value, and thus renders a contradiction.

Now, consider the case of infinite N for which $\mathbf{P}_n = \mathbf{P}$ (see Section 2) and assuming that $T_n = T$ for all $n = 1, 2, \ldots$, where T represents the duration for each epoch. The value function $J_n(i, \mathbf{C}_n)$ is then easily shown to be independent of n such that for any state i and any capacity vector $\mathbf{C}$ it is given by

$$J(i, \mathbf{C}) = \max_{\mathbf{x}_k} \sum_{k=0}^{\infty} e^{-\beta k} \sum_{j=1}^{U} \mathbf{P}_{ij}^{k} H(j, \mathbf{C} + \sum_{\ell=0}^{k-1} \mathbf{x}_\ell, \mathbf{x}_k) \qquad (17)$$

where

$$H(i, \mathbf{C}, \mathbf{x}) \triangleq R(i, \mathbf{C})T - \mathbf{K} \cdot \mathbf{C}T - \mathbf{I} \cdot \mathbf{x}^{+}(i) - \mathbf{D} \cdot \mathbf{x}^{-}(i).$$

From standard dynamic programming arguments (see, e.g., [21, 6]), we know that $J(i, \mathbf{C})$ satisfies the Bellman equation

$$J(i, \mathbf{C}) = \max_{\mathbf{x}} \left\{ H(i, \mathbf{C}, \mathbf{x}) + e^{-\beta} \sum_{j=1}^{U} \mathbf{P}_{ij} J(i, \mathbf{C} + \mathbf{x}) \right\}. \quad (18)$$

Although this is a typical stochastic dynamic program, the multidimensional aspects of the problem cause it to be very difficult to solve. When we consider $\mathbf{C}$ taking on only integer values, the problem is equivalent to solving the following linear program

$$\min \sum_{i=1}^{U} \sum_{\mathbf{C}} \boldsymbol{\gamma}(i) \prod_{j=1}^{J} \eta_j^{C_j} J(i, \mathbf{C})$$

$$\text{s.t. } J(i, \mathbf{C}) \geq \max_{\mathbf{x}} \left\{ H(i, \mathbf{C}, \mathbf{x}) + e^{-\beta} \sum_{j=1}^{U} \mathbf{P}_{ij} J(j, \mathbf{C} + \mathbf{x}) \right\} \quad (19)$$

for some $\beta, \eta_1, \cdots, \eta_J \in (0, 1)$. Known techniques for infinite-dimensional linear programs [1] can be employed to solve or approximate this problem. Another alternative is to truncate the linear program and use ordinary methods to solve the truncated version. Most importantly, however, the special structure of the objective function enables us to show that the ergodic version of the above optimization problem has finite solutions and the optimal solution can be obtained as the limit of the discounted problem.

THEOREM 3.1. *There exists a constant ψ and a function $J_E(i, \mathbf{C})$ that satisfies the following relationship*

$$\psi + J_E(i, \mathbf{C}) = \max_{\mathbf{x}} \left\{ H(i, \mathbf{C}, \mathbf{x}) + \sum_{j=1}^{U} \mathbf{P}_{ij} J_E(i, \mathbf{C} + \mathbf{x}) \right\}. \quad (20)$$

PROOF. Let $J_\beta(i, \mathbf{C})$ denote the solution of (18), emphasizing the discount factor β. Now fix some state, say $(1, \mathbf{0})$, and define

$$G_\beta(i, \mathbf{C}) \triangleq J_\beta(i, \mathbf{C}) - J_\beta(1, \mathbf{0}), \quad \forall i = 1, 2, \cdots, U.$$

From our assumptions on $\mathbf{P}$, we know that $G_\beta(i, \mathbf{C})$ satisfies

$$\begin{aligned} & (1 - e^{-\beta}) J_\beta(1, \mathbf{0}) + G_\beta(i, \mathbf{C}) \\ = \; & \max_{\mathbf{x}} \left\{ H(i, \mathbf{C}, \mathbf{x}) + e^{-\beta} \sum_{j=1}^{U} \mathbf{P}_{ij} J_\beta(j, \mathbf{C} + \mathbf{x}) \right\}. \quad (21) \end{aligned}$$

It then follows from (17) that $J_\beta(i, \mathbf{C})$ increases with decreasing β. To see this, suppose $\mathbf{x}^*$ is the optimal vector for $\beta = \beta_1$, and thus we have for $\beta_2 < \beta_1$

$$\begin{aligned} & \sum_{n=0}^{\infty} e^{-\beta_2 n} \sum_{j=1}^{U} \mathbf{P}_{ij}^{n} H(j, \mathbf{C} + \sum_{\ell=0}^{n-1} \mathbf{x}_\ell^*, \mathbf{x}_n^*) \\ \geq \; & \sum_{n=0}^{\infty} e^{-\beta_1 n} \sum_{j=1}^{U} \mathbf{P}_{ij}^{n} H(j, \mathbf{C} + \sum_{\ell=0}^{n-1} \mathbf{x}_\ell^*, \mathbf{x}_n^*), \end{aligned}$$

which implies $J_{\beta_2}(i, \mathbf{C}) \geq J_{\beta_1}(i, \mathbf{C})$. This monotonicity property guarantees the convergence of $J_\beta(i, \mathbf{C})$ as $\beta \to 0$. Then equation (21) renders (20) as long as $(1 - e^{-\beta}) J_\beta(1, \mathbf{0})$ does not go to infinity as $\beta \to 0$. Meanwhile, letting R^* denote the total possible revenue collected in each epoch, we obviously have for any $i, \mathbf{C}$

$$\begin{aligned} & \sum_{n=0}^{\infty} e^{-\beta n} \sum_{j=1}^{U} \mathbf{P}_{ij}^{n} H(j, \mathbf{C} + \sum_{\ell=0}^{n-1} \mathbf{x}_\ell, \mathbf{x}_n) \\ \leq \; & \sum_{n=0}^{\infty} e^{-\beta n} R^* = R^* (1 - e^{-\beta})^{-1}. \end{aligned}$$

Hence, $(1 - e^{-\beta}) J_\beta(1, \mathbf{0})$ can be bounded from above by R^* as $\beta \to 0$, and the theorem statement follows. □

3.2 Optimality of Solution

We next establish that the capacity planning policy $\mathbb{C}^*$ obtained from the results in the previous section is asymptotically optimal with respect to the original, general stochastic loss network formulation of Section 2 (as opposed to the limiting regime of this formulation). To do so, consider a sequence of stochastic loss networks, indexed by $k = 1, 2, \ldots$, in which the traffic intensity vector is scaled by k, namely the traffic intensity for route $r \in \mathcal{R}$ in the kth loss network is $k\nu_r$. For any feasible capacity planning policy $\mathbb{Y}_{(k)}$ that employs the capacity vector $\mathbf{Y}_{n,(k)} = (\mathbf{Y}_{n,(k)}(1), \ldots, \mathbf{Y}_{n,(k)}(U))$ throughout epoch n in the kth stochastic loss network, define $\mathbf{x}_{n,(k)} \triangleq \mathbf{Y}_{n+1,(k)} - \mathbf{Y}_{n,(k)}$ and let $J_{(k)}^{\mathbb{Y}(k)}$ be the corresponding discounted total profit given by

$$J_{(k)}^{\mathbb{Y}(k)} = \mathbb{E}_{\mathbb{Y}_{(k)}} \left[\sum_{n=0}^{N-1} \sum_{i=1}^{U} e^{-\beta n} \left(\prod_{\ell=0}^{n-1} \mathbf{P}_\ell \right) H_{n,(k)}^{\mathbb{Y}(k)}(i, \mathbf{Y}_{n,(k)}, \mathbf{x}_{n,(k)}) \right],$$

where $\mathbb{E}_{\mathbb{Y}_{(k)}}$ denotes expectation taken under the policy $\mathbb{Y}_{(k)}$.

DEFINITION 3.1. *A capacity planning policy $\mathbb{C}_{(k)}^*$ is called* asymptotically optimal *if for any feasible policy $\mathbb{Y}_{(k)}$, we have*

$$\limsup \frac{J_{(k)}^{\mathbb{Y}(k)}}{J_{(k)}^{\mathbb{C}_{(k)}^*}} \;\leq\; 1, \qquad \textit{as } k \to \infty. \qquad (22)$$

This definition indicates that the profit $J_{(k)}^{\mathbb{C}_{(k)}^*}$ is the best profit one can achieve asymptotically (i.e., as the system grows large) and that this asymptotically maximal profit is achieved by the sequence of capacity planning policies $\{\mathbb{C}_{(k)}^*\}$. Note that this definition of asymptotic optimality is consistent with what has been established in the research literature, e.g., refer to [20, 4, 2].

Now we present our main result on asymptotic optimality.

THEOREM 3.2. *The capacity planning policy $\mathbb{C}^*$ is asymptotically optimal.*

PROOF. For any sequence of policies $\mathbb{Y}_{(k_\ell)}$, let $J_{(k_\ell)}^{\mathbb{Y}(k_\ell)}$ be a subsequence of $J_{(k)}^{\mathbb{Y}(k)}$ where the corresponding capacity vector $\mathbf{Y}_{n,(k_\ell)}$ is such that, as $k_\ell \to \infty$, $\mathbf{Y}_{n,(k_\ell)}(i)/k_\ell$ converges componentwise. The proof proceeds based on the two different cases of convergence.

Let us first consider the case where $\mathbf{Y}_{n,(k_\ell)}(i)/k_\ell$ converges componentwise to a finite real number for any epoch n and $i = 1, 2, \cdots, U$, namely there exists a $\mathbf{Y}_n(i)$ such that

$$\lim_{\ell \to \infty} \mathbf{Y}_{n,(k_\ell)}(i)/k_\ell \;=\; \mathbf{Y}_n(i).$$

It is well-known (cf. [23, 15]) that there exists a vector $\mathbf{B}^{\mathbb{Y}} = (\mathbf{B}_1^{\mathbb{Y}}, \ldots, \mathbf{B}_J^{\mathbb{Y}}) \in [0,1]^J$ satisfying the conditions (13),(14), where the blocking probability $\mathbf{B}_{n,j,(k_\ell)}^{\mathbb{Y}(k_\ell)}(i)$ for link j of each loss network k_ℓ converges to $\mathbf{B}_{n,j}^{\mathbb{Y}}(i)$ for all $j = 1, \ldots, J$. From the continuity of the objective function, we have

$$\lim_{\ell\to\infty} \frac{J_{(k_\ell)}^{\mathbb{Y}(k_\ell)}}{k_\ell} \;=\; \mathbb{E}_{\mathbb{Y}}\left[\sum_{n=0}^{N-1}\sum_{i=1}^{U} e^{-\beta n}\left(\prod_{\ell=0}^{n-1}\mathbf{P}_\ell\right) H_n^{\mathbb{Y}}(i, \mathbf{Y}_n, \mathbf{x}_n)\right],$$

where $H_n^{\mathbb{Y}}$ is evaluated with respect to $\mathbf{Y}_n$ and $\mathbf{x}_n$, and, by the definition of $\mathbb{C}^*$, we conclude

$$\lim_{\ell\to\infty} \frac{J_{(k_\ell)}^{\mathbb{Y}(k_\ell)}}{J_{(k_\ell)}^{\mathbb{C}^*_{(k_\ell)}}} \;=\; \lim_{\ell\to\infty} \frac{J_{(k_\ell)}^{\mathbb{Y}(k_\ell)}/k_\ell}{J_{(k_\ell)}^{\mathbb{C}^*_{(k_\ell)}}/k_\ell} \;\le\; 1,$$

which establishes the desired result for this case.

Lastly, let us consider the case where there exists at least one n_0 such that

$$\lim_{\ell\to\infty} \mathbf{Y}_{n_0,(k_\ell)}(i)/k_\ell \;=\; \boldsymbol{\infty}.$$

This implies $\mathbf{B}_{n_0}^{\mathbb{Y}(k_\ell)}(i) = \mathbf{0}$, and thus from (8) the revenue function is linear subject to the traffic intensity vector. It follows that $H_{n_0,(k_\ell)}^{\mathbb{Y}(k_\ell)}(i, \mathbf{Y}_{n_0,(k_\ell)}, \mathbf{x}_{n_0,(k_\ell)})/k_\ell \to -\infty$ since the numerator is bounded from above by $k_\ell \mathbf{w} \cdot \boldsymbol{\nu}_{n_0}(i) T_{n_0} - \mathbf{K} \cdot \mathbf{Y}_{n_0,(k_\ell)}(i) T_{n_0}$, and therefore we have

$$\frac{J_{(k_\ell)}^{\mathbb{Y}(k_\ell)}}{J_{(k_\ell)}^{\mathbb{C}^*_{(k_\ell)}}} \;=\; \frac{J_{(k_\ell)}^{\mathbb{Y}(k_\ell)}/k_\ell}{J_{(k_\ell)}^{\mathbb{C}^*_{(k_\ell)}}/k_\ell} \;\to\; -\infty, \qquad \text{as } \ell \to \infty,$$

which satisfies (22) and completes the proof. □

The above formulation and Theorem 3.2 apply to the finite-horizon version of our stochastic dynamic programming problem in the limiting regime. The corresponding infinite-horizon result follows from the arguments used to establish Theorems 3.1 and 3.2 under the appropriate limits. In the interest of space, we omit these details.

4. GENERAL OPTIMALITY

The previous section provides a solution to the stochastic capacity planning problem (11),(12) in the limiting regime of our original formulation over a time horizon of N epochs, for both finite and infinite N, and establishes that this solution is asymptotically optimal with respect to the original formulation. In this section, we return to the general stochastic capacity planning optimization problem (11),(12). However, since alternative objective functions of general stochastic loss networks may be of interest both in theory and in practice, we broaden our scope even further to consider a related stochastic optimization problem in a very general setting (not even restricted to stochastic loss networks), for which the original formulation is a special case. We prove that optimal solutions of finite-horizon and infinite-horizon versions of this problem have a threshold structure, similar to that determined in the previous section. Then we discuss some of the implications of these results, including how our results can be used to efficiently compute the capacity (threshold) values of the optimal policy. Finally, we briefly discuss two specific applications that exploit our general stochastic optimization results, the first based on the Erlang fixed-point approximation within the context of a network setting and the second based on a related application in manufacturing systems.

4.1 Stochastic Dynamic Program

For any state i of the Markov-modulated process $\mathbf{P}_n$ starting at any epoch n and any nonnegative capacity vectors $\mathbf{C}$ and $\mathbf{x}$, define

$$H_n^N(i, \mathbf{C}, \mathbf{x}) \triangleq R_n(i, \mathbf{C}) T_n - \mathbf{K} \cdot \mathbf{C} T_n - \mathbf{I} \cdot \mathbf{x}^+ - \mathbf{D} \cdot \mathbf{x}^-, \quad (23)$$

where N is the total number of epochs. (We include N in our notation because we will consider the limit as $N \to \infty$ in Theorem 4.2.) Let $J_n^N(i, \mathbf{C}_n)$ be the value function for maximizing the expected discounted profit of the stochastic optimization problem of interest over the time horizon from epoch n to epoch $N-1$ starting in state i and epoch n with capacities $\mathbf{C}_n$ and traffic intensities $\boldsymbol{\nu}_n$. Then we can formulate our general stochastic optimization problem in terms of the Bellman optimality equations

$$J_n^N(i, \mathbf{C}_n) \;=\; \max_{\mathbf{x}_n} P_n^N(i, \mathbf{C}_n, \mathbf{x}_n), \quad (24)$$

$$P_n^N(i, \mathbf{C}_n, \mathbf{x}_n) \;=\; e^{-\beta} \sum_{j=1}^{U} \mathbf{P}_{n,ij} J_{n+1}^N(j, \mathbf{C}_n + \mathbf{x}_n) + e^{-\beta} H_n^N(i, \mathbf{C}_n, \mathbf{x}_n), \quad (25)$$

where we assume $J_N^N(i, \mathbf{C}_N) = 0$.

The general setting for the stochastic optimization problems of interest in this section are based on the following assumptions.

ASSUMPTION 4.1. *For each fixed $i = 1, \ldots, U$ and every epoch n, assume $R_n(i, \mathbf{C})$ is concave in $\mathbf{C}$ and that*

$$\lim_{\mathbf{C}\to\infty} \frac{R_n(i, \mathbf{C})}{\mathbf{K} \cdot \mathbf{C}} \;=\; 0,$$

where convergence is componentwise.

ASSUMPTION 4.2. *Assume the values of $R_n(i, \mathbf{C})$, $\mathbf{K}$, $\mathbf{I}$ and $\mathbf{D}$ are such that the smallest capacity vector $\mathbf{Y} = \mathbf{C} + \mathbf{x}$ maximizing $e^{-\beta} H_n^N(i, \mathbf{C}, \mathbf{x})$ is finite and nonnegative, and thus $e^{-\beta} H_n^N(i, \mathbf{C}, \mathbf{x})$ is nonnegative for every $i = 1, \ldots, U$ in any epoch n.*

Assumption 4.1 is a reflection of the typical reality in practice that the marginal effect of increasing capacity on revenue does not increase and eventually diminishes. The assumptions in 4.1 are also related to common assumptions in much of the relevant economic theory on utility functions. Assumption 4.2 is a reflection of the typical reality in practice that the revenue functions and cost functions must allow for a profitable solution, because otherwise the optimal solution is to completely avoid the capacity planning opportunity in order to prevent financial ruin. The assumptions in 4.2 further exclude the case where the revenue functions take on the special form of linear combinations of $\mathbf{I} \cdot \mathbf{x}^+$ and $\mathbf{D} \cdot \mathbf{x}^-$, in which case the problem can be easily reduced to the form considered in Section 3.

We now can present our first main result of this section for the finite-horizon version of the optimization problem, where all vector limits are with respect to componentwise convergence.

THEOREM 4.1. *Suppose Assumptions 4.1 and 4.2 hold. Then, for each fixed i and all $n = 0, \ldots, N-1 < \infty$, there exists a finite capacity vector that realizes the global optimal solution of problem (24),(25) starting at epoch n and this optimal solution is the capacity planning policy $\mathbb{X}_{n,N}^*$ that employs the capacity vector $\mathbf{X}_{n,N}^* = (\mathbf{X}_{n,N}^*(1), \ldots, \mathbf{X}_{n,N}^*(U))$ where $\mathbf{X}_{n,N}^*(i)$ is the smallest capacity vector $\mathbf{Y} = \mathbf{C} + \mathbf{x}$ that maximizes $P_n^N(i, \mathbf{C}, \mathbf{x})$, $i = 1, \ldots, U$.*

PROOF. Obviously, Assumption 4.1 implies that $e^{-\beta} H_n^N(i, \mathbf{C}, \mathbf{x})$ is concave in $\mathbf{Y} = \mathbf{C} + \mathbf{x}$ and that

$$\lim_{\mathbf{Y}=\mathbf{C}+\mathbf{x}\to\infty} e^{-\beta} H_n^N(i, \mathbf{C}, \mathbf{x}) \; = \; -\infty$$

for each fixed i and any $n = 0, \ldots, N-1$.

The proof proceeds by induction where we first consider the basis step $n = N - 1$. From (25) and the corresponding properties of $e^{-\beta} H_n^N(i, \mathbf{C}, \mathbf{x})$ together with $J_{n+1}^N(i, \mathbf{C}) = 0$, it follows that $P_n^N(i, \mathbf{C}, \mathbf{x})$ is concave in $\mathbf{Y} = \mathbf{C} + \mathbf{x}$ and that

$$\lim_{\mathbf{Y}=\mathbf{C}+\mathbf{x}\to\infty} P_n^N(i, \mathbf{C}, \mathbf{x}) \; = \; -\infty.$$

These properties and (24) imply that $J_n^N(i, \mathbf{C})$ is concave in $\mathbf{C}$ and that

$$J_n^N(i, \mathbf{C}) \; \geq \; J_{n+1}^N(i, \mathbf{C}) \; \geq \; 0. \tag{26}$$

Since $P_n^N(i, \mathbf{C}, \mathbf{x})$ is a positive linear combination of the quantities $e^{-\beta} H_n^N(i, \mathbf{C}, \mathbf{x})$ and the $J_{n+1}^N(i, \mathbf{C})$, it follows from (26) and Assumption 4.2 that

$$P_{n-1}^N(i, \mathbf{C}, \mathbf{x}) \; \geq \; P_n^N(i, \mathbf{C}, \mathbf{x}) \; \geq \; 0.$$

Hence, there exists a finite capacity vector that realizes the global maximum of $P_n^N(i, \mathbf{C}, \mathbf{x})$, and the optimal policy for the problem starting at epoch $n = N - 1$ is $\mathbb{X}_{n,N}^*$ where $\mathbf{X}_{n,N}^*(i)$ is the smallest capacity vector $\mathbf{Y} = \mathbf{C} + \mathbf{x}$ that maximizes $P_n^N(i, \mathbf{C}, \mathbf{x})$ since $J_{n+1}^N(i, \mathbf{C}) = 0$.

Next, as part of the induction step, suppose the above statements are true for $n + 1$ and consider the problem starting at epoch n. Then the four induction step properties

$$\begin{aligned}
&\text{(i)} && J_{n+1}^N(i, \mathbf{C}) \;\text{ is concave in }\; \mathbf{C}, \\
&\text{(ii)} && P_n^N(i, \mathbf{C}, \mathbf{x}) \; \geq \; P_{n+1}^N(i, \mathbf{C}, \mathbf{x}) \; \geq \; 0, \\
&\text{(iii)} && P_{n+1}^N(i, \mathbf{C}, \mathbf{x}) \;\text{ is concave in }\; \mathbf{Y} = \mathbf{C} + \mathbf{x}, \\
&\text{(iv)} && \lim_{\mathbf{Y}=\mathbf{C}+\mathbf{x}\to\infty} P_{n+1}^N(i, \mathbf{C}, \mathbf{x}) \; = \; -\infty
\end{aligned} \tag{27}$$

together with (25) imply that $P_n^N(i, \mathbf{C}, \mathbf{x})$ is concave in $\mathbf{Y} = \mathbf{C} + \mathbf{x}$ and that

$$\lim_{\mathbf{Y}=\mathbf{C}+\mathbf{x}\to\infty} P_n^N(i, \mathbf{C}, \mathbf{x}) \; = \; -\infty.$$

It follows from these properties and (24) that $J_n^N(i, \mathbf{C})$ is concave in $\mathbf{C}$. Now observe that

$$J_n^N(i, \mathbf{C}) - J_{n+1}^N(i, \mathbf{C}) \; \geq \; \max_{\mathbf{x}} \{P_n^N(i, \mathbf{C}, \mathbf{x}) - P_{n+1}^N(i, \mathbf{C}, \mathbf{x})\},$$

which together with the induction step property (27) imply that

$$J_n^N(i, \mathbf{C}) \; \geq \; J_{n+1}^N(i, \mathbf{C}) \; \geq \; 0.$$

It follows from this property and Assumption 4.2 that

$$P_{n-1}^N(i, \mathbf{C}, \mathbf{x}) \; \geq \; P_n^N(i, \mathbf{C}, \mathbf{x}) \; \geq \; 0,$$

since $P_n^N(i, \mathbf{C}, \mathbf{x})$ is a positive linear combination of the quantities $e^{-\beta} H_n^N(i, \mathbf{C}, \mathbf{x})$ and the $J_{n+1}^N(i, \mathbf{C})$. Hence, there exists a finite capacity vector that realizes the global maximum of $P_n^N(i, \mathbf{C}, \mathbf{x})$, and the optimal policy for the problem starting at epoch n is $\mathbb{X}_{n,N}^*$ where $\mathbf{X}_{n,N}^*(i)$ is the smallest capacity vector $\mathbf{Y} = \mathbf{C} + \mathbf{x}$ that maximizes $P_n^N(i, \mathbf{C}, \mathbf{x})$, which completes the proof. □

We next establish our second main result of this section for the infinite-horizon version of the stochastic optimization problem, where all vector limits are with respect to componentwise convergence.

THEOREM 4.2. *Suppose Assumptions 4.1 and 4.2 hold. Then, letting $N \to \infty$, the sequences $\{J_0^N\}$ and $\{P_0^N\}$ converge pointwise to limits J_0^∞ and P_0^∞, respectively, where for all $i = 1, \ldots, U$ $J_0^\infty(i, \mathbf{C})$ is concave in $\mathbf{C}$,*

$$\lim_{\mathbf{C}\to\infty} J_0^\infty(i, \mathbf{C}) \; = \; -\infty,$$

$P_0^\infty(i, \mathbf{C}, \mathbf{x})$ is concave in $\mathbf{Y} = \mathbf{C} + \mathbf{x}$, and

$$\lim_{\mathbf{Y}=\mathbf{C}+\mathbf{x}\to\infty} P_0^\infty(i, \mathbf{C}, \mathbf{x}) \; = \; -\infty.$$

Hence, there exists a finite capacity vector that realizes the global optimal solution of the infinite-horizon problem (24),(25) and this optimal solution is the capacity planning policy $\mathbb{X}^*$ that employs the capacity vector $\mathbf{X}^*(i)$ whenever the system is in state i where $\mathbf{X}^*(i)$ is the smallest capacity vector $\mathbf{Y} = \mathbf{C} + \mathbf{x}$ that maximizes $P_0^\infty(i, \mathbf{C}, \mathbf{x})$, $i = 1, \ldots, U$.

PROOF. The statements concerning the limits J_0^∞ and P_0^∞ and the properties of $J_0^\infty(i, \mathbf{C})$ and $P_0^\infty(i, \mathbf{C}, \mathbf{x})$ follow directly from Theorem 4.1 provided there exists a finite nondecreasing function $\mathcal{B}_J(\mathbf{C})$ such that $J_0^N(i, \mathbf{C}) \leq \mathcal{B}_J(\mathbf{C})$ for all N and all $i = 1, \ldots, U$. From the induction step of Theorem 4.1, we know there exists a finite nondecreasing function $\mathcal{B}_P(\mathbf{C}, \mathbf{x})$ such that $P_0^N(i, \mathbf{C}, \mathbf{x}) \leq \mathcal{B}_P(\mathbf{C}, \mathbf{x})$ for all N and all $i = 1, \ldots, U$. Replacing the right-hand side of (24) with $\mathcal{B}_P(\mathbf{C}, \mathbf{x})$ and applying the induction arguments of Theorem 4.1 to this revised optimization problem reveals the existence of a finite nondecreasing function $\mathcal{B}_J(\mathbf{C})$ such that $J_0^N(i, \mathbf{C}) \leq \mathcal{B}_J(\mathbf{C})$ for all N and all $i = 1, \ldots, U$.

We next note that

$$\mathbf{X}_{0,\infty}^*(i) \; = \; \lim_{N\to\infty} \mathbf{X}_{0,N}^*(i)$$

and there exists a constant N_0 such that $\mathbf{X}_{0,N}^*(i) = \mathbf{X}_{0,\infty}^*(i)$ for all $N \geq N_0$. Then, if $\mathbf{C} \leq \mathbf{X}_{0,\infty}^*(i)$,

$$J_0^N(i, \mathbf{C}) \; = \; P_0^N(i, \mathbf{C}, \mathbf{X}_{0,\infty}^*(i) - \mathbf{C})$$

and otherwise

$$J_0^N(i, \mathbf{C}) \; = \; P_0^N(i, \mathbf{C}, \mathbf{0}),$$

for all $N \geq N_0$. Letting $N \to \infty$ and substituting

$$P_0^\infty(i, \mathbf{C}, \mathbf{X}^*(i) - \mathbf{C}) \; = \; P_0^\infty(i, \mathbf{C}, \mathbf{X}_{0,\infty}^*(i) - \mathbf{C})$$

and $\mathbf{X}^*(i) \leq \mathbf{X}_{0,\infty}^*(i)$ yields

$$J_0^\infty(i, \mathbf{C}) \; = \; P_0^\infty(i, \mathbf{C}, \mathbf{X}^*(i) - \mathbf{C})$$

if $\mathbf{C} \leq \mathbf{X}^*(i)$, and otherwise

$$J_0^\infty(i, \mathbf{C}) \; = \; P_0^\infty(i, \mathbf{C}, \mathbf{0}).$$

It follows from the definition of $\mathbf{X}^*(i)$ that

$$J_0^\infty(i, \mathbf{C}) \; = \; \max_{\mathbf{x}} P_0^\infty(i, \mathbf{C}, \mathbf{x})$$

and the limit $P_0^\infty(i, \mathbf{C}, \mathbf{x})$ is given by (25) in terms of the $J_0^\infty(i, \mathbf{C} + \mathbf{x})$ and $e^{-\beta} H_0^\infty(i, \mathbf{C}, \mathbf{x})$. This together with the statements concerning the limit J_0^∞, the properties of $J_0^\infty(i, \mathbf{C})$ and well-known results from the infinite-horizon stochastic dynamic programming theory (see, e.g., [21, 6]) complete the proof. □

4.2 Discussion of Optimal Solutions

In Theorems 4.1 and 4.2 we establish that the optimal solution of our capacity planning problem is an optimal threshold-based policy in the case of finite and infinite time horizons, respectively. This structural property is important because threshold policies are easy to implement in practice. For each state, we need only solve a sequence of convex programs to obtain the optimal capacity threshold values, as discussed in more detail below, and then set the capacity of every link to match these optimal thresholds. Furthermore, the threshold-based structural property, together with related properties of the value functions such as convexity (concavity) and semimodularity (refer to Theorems 4.1 and 4.2), play important roles in developing the complete set of structural properties for the optimal solutions and for the general optimization problems themselves.

To compute the capacity vector $\mathbf{X}^*_{n,N} = (\mathbf{X}^*_{n,N}(1), \dots, \mathbf{X}^*_{n,N}(U))$ deployed throughout each epoch $n = 0, \dots, N-1$ under the optimal capacity planning policy $\mathbb{X}^*_{n,N}$ for the finite-horizon problem, we start with the problem at epoch $n = N-1$ in any state $i = 1, \dots, U$. From Theorem 4.1 we have that the vector $\mathbf{X}^*_{N-1,N}(i)$ is the solution of a convex (concave) program which can be computed in an efficient manner using known methods in convex optimization; e.g., refer to [3, 5]. Then we recursively continue in this manner to solve the problem for each epoch until we obtain the set of capacity vectors comprising the optimal capacity planning policy $\mathbb{X}^*_{n,N}$ for all $n = 0, \dots, N-1$. Analogously, from Theorem 4.2, the optimal capacity vector $\mathbf{X}^*(i)$ for each state $i = 1, \dots, U$ in the infinite-horizon problem can be obtained by solving a convex (concave) program. Upon computing these solutions for all $i = 1, \dots, U$, we obtain the set of capacity vectors comprising the optimal capacity planning policy $\mathbb{X}^*$.

4.3 Applications

Lastly, let us turn to briefly consider two specific examples for which our general stochastic optimization results above can be applied. The first example is based on the Erlang fixed-point approximation in a communications network setting where the expected revenue rate $R_n(i, \mathbf{C})$ for a network service provider given a traffic intensity vector $\boldsymbol{\nu}_n(i)$ and a nonnegative capacity vector $\mathbf{C}$ is as expressed in (8). Although this function $R_n(i, \mathbf{C})$ obtained from the Erlang fixed-point approximation is generally not concave in $\mathbf{C}$, there are regions in which the function $R_n(i, \mathbf{C})$ satisfies the concavity part of Assumption 4.1 [16]. Furthermore, generally speaking, it is possible to define revenue functions $R_n(i, \mathbf{C})$ based on the Erlang fixed-point approximation that are concave in $\mathbf{C}$ by using alternative formulations from that in (8). It is also well-known that the Erlang loss formula for a single link of capacity C, as provided in (7), is convex in C for every traffic intensity ν [14]. In all such cases where $R_n(i, \mathbf{C})$ is concave with respect to $\mathbf{C}$, it is important to identify values of costs $\mathbf{K}$, $\mathbf{I}$ and $\mathbf{D}$ so that Assumptions 4.1 and 4.2 are satisfied, because otherwise the optimal solution would be for the service provider to always maintain a capacity vector of $\mathbf{0}$ and not serve any traffic. Once Assumptions 4.1 and 4.2 have been verified, the results of Theorems 4.1 and 4.2 provide the optimal policy and the corresponding convex programming solutions for the capacity vectors of this optimal policy that should be followed by the network service provider in the case of finite-horizon and infinite-horizon versions of the capacity planning optimization problem, respectively.

The second example is a multi-product production inventory system, where the capacities of different classes are the base-stock levels for different products [28]. Let us denote these levels as

$$\mathbf{S}_n = (S_{n,1}, S_{n,2}, \cdots, S_{n,J}).$$

Multi-class inventory system workloads are classified by the different combinations of products they require, analogous to the link requirements of routes. Upon treating the replenishment lead time as the random duration of our model, we can see that the multi-product production-inventory system fits quite well within our model formulation. For a related study of such inventory systems, see, e.g., [26]. In addition, a more popular performance metric that has been extensively studied in the operations management literature is the average amount of backlogged units in a system where backlogging is allowed. The backlogs are characterized by complex combinations of maximum and addition operators of the system statistics and base-stock levels. However, it has been shown that under mild conditions, the average backlogs are convex functions of $\mathbf{S}_n$; see, e.g., [19]. Hence, we can define a concave profit function with respect to $\mathbf{S}_n$. Upon combining this with inventory costs and the costs incurred for increasing and decreasing the base-stock levels, we can use the formulation of (24) and (25) to determine the optimal base-stock level $\mathbf{S}_n$ for each epoch n, where it can be easily verified that the Assumptions 4.1 and 4.2 are satisfied.

	Network 1	Network 2
$\mathbf{K}$	(15860, 10660, 21060)	(15860, 10660, 21060, 10660, 15860, 15860)
$\mathbf{w}$	$(60, 80) \times 10^3$	$(192, 252, 168, 115.5) \times 10^3$
$\mathbf{I}$	$(1000, 1500, 750)$	$(1000, 800, 600, 800, 750, 900)$
$\boldsymbol{\mu}$	(1, 1)	(1, 1, 1, 1)
T_n	65	65
β	$\log_e(0.8)$	$\log_e(0.8)$

Table 1: Base model parameters for Network 1 and Network 2.

5. NUMERICAL EXPERIMENTS

In this section, we present a representative sample of numerical experiments with some of our optimal capacity planning solutions. A stochastic dynamic program will be solved based on the Erlang fixed-point approximation, either in the limiting regime or in its standard form, and the optimal capacity allocations for time-varying workloads will be calculated and discussed.

We consider two stochastic loss networks.

- Network 1: A 3-link network with capacities C_1, C_2 and C_3 supporting calls on routes $r_1, r_2 \in \mathcal{R}$ according to the capacity requirement matrix

  $$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix},$$

  where $A_{rj} = 1$ if route r requires a unit of capacity on link j as defined in Section 2.

- Network 2: A 6-link network with capacities C_i, $i \in \{1, \dots, 6\}$, supporting calls on routes r_j, $j \in \{1, \dots, 4\}$, according to the capacity requirement matrix

  $$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

The base settings for all other model parameters are provided in Table 1.

For Network 1, we study the trajectories of the optimal capacity vectors corresponding to the stochastic optimization in (11),(12) under the following 3 traffic intensity demand profiles for a time horizon of $N = 5$ epochs:

- Profile 1: (80, 90), (75, 60), (60, 75), (55, 45), (40, 45)

- Profile 2: (35, 45), (55, 45), (60, 70), (70, 80), (80, 90)

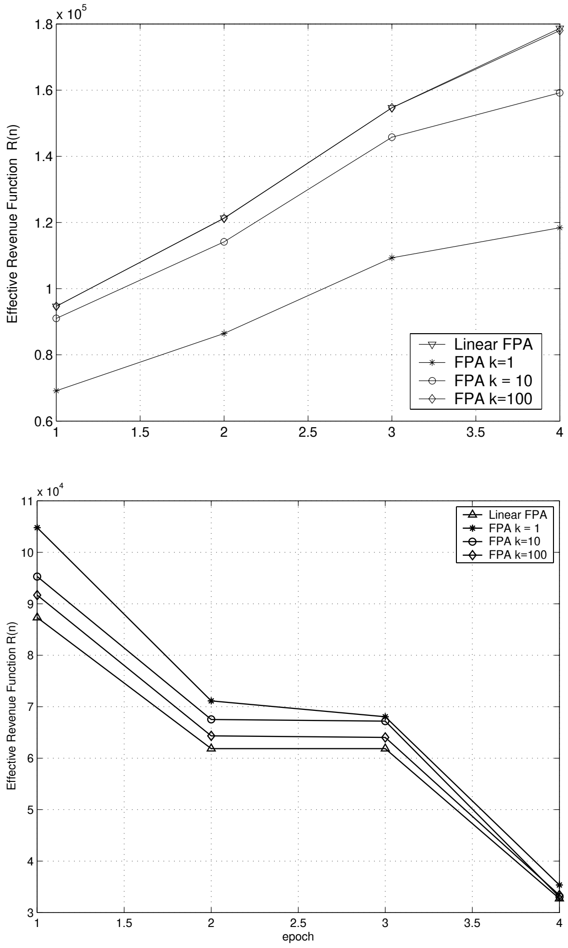

Figure 1: Optimal Revenues $R(n)$ for (top) demand Profile 2 on Network 1, and (bottom) demand Profile 1 on Network 2.

- Profile 3: (75, 60), (60, 55), (75, 60), (60, 55), (75, 60)

The corresponding demand profiles for Network 2 are given by

- Profile 1: (80, 90, 70, 65), (70, 80, 60, 65), (60, 70, 55, 50), (55, 45, 55, 40), (35, 45, 50, 40)
- Profile 2: (35, 45, 50, 40), (55, 45, 55, 40), (60, 70, 55, 50), (70, 80, 60, 65), (80, 90, 70, 65)
- Profile 3: (35, 45, 50, 40), (70, 80, 60, 65), (35, 45, 50, 40), (70, 80, 60, 65), (35, 45, 50, 40)

The optimal solutions are obtained by dynamic programming over a quantized set of possible capacity vectors. The complexity of the dynamic program scales as $\Omega(\text{NumLevel}^6 \times \xi^3)$ where NumLevel is the number of quantization steps and ξ is the complexity of the operation to calculate the Erlang fixed point. In our experiments, the Erlang fixed point calculation using the linear approximation from Section 3, i.e., in the limiting regime, yielded results that were two orders of magnitude faster than solving the Erlang fixed-point equations (6). This leads to an overall potential savings of 10^6 in the computational complexity of using the linear approximation for the case of Network 1 and a corresponding savings of 10^{12} for the case of Network 2.

Our first set of experiments compares the results from the optimal solutions for the Erlang fixed-point approximation in the limiting regime and for the Erlang fixed-point approximation expressed in (6) under different scalings of the traffic intensities. More specifically, in Figure 1, we demonstrate the strong accuracy of the solution based on the linear approximation with respect to the solution based on the Erlang fixed-point approximation under the scaling factors $k = 1, 10, 100$. Our results for Profile 2 on Network 1 and Profile 1 on Network 2 are presented where, in all cases, the expected revenues are calculated using the corresponding optimal solutions. It is interesting to observe that, even for relatively small values of k, the expected revenue results from the linear approximation hew very closely to those from the Erlang fixed-point approximations. The same observations can be made for the optimal capacities in the corresponding stochastic optimization problems. Theoretically, the convergence is on the order of $1/\sqrt{k}$, which is not a very fast rate, but this result is obtained without making any use of the correlation information that is quite rich in the stochastic systems under consideration. We suspect that the real convergence rate will be much higher, which, of course, warrants further study and a detailed analysis of the asymptotic behavior of the stochastic loss networks.

Our next collection of experiments examines the key characteristics and trends of the optimal capacity thresholds from the solution of the stochastic capacity planning optimization problem in terms of the different demand profiles, networks and model parameters. Based on the results in Figure 1, we shall focus on the optimal solutions using the Erlang fixed-point approximation in the limiting regime. Figure 2 presents a representative sample of our results for Profiles 1 – 3 in Network 2 under the base model parameter settings in Table 1. The leftmost plots of Figures 3 – 5 provide the corresponding base case results for Profiles 1 – 3, respectively, in Network 1. For comparison, the rightmost plots in these figures present the corresponding results for the case where the call duration rates $\boldsymbol{\mu} = (1, 1)$ are changed to $\boldsymbol{\mu}' = (1, 2)$, i.e., the mean duration for r_1 is twice that of r_2, and the center plots provide the corresponding results for the case where the capacity costs $\mathbf{K} = (15860, 10660, 21060)$ are changed to $\mathbf{K}' = (31720, 10660, 21060)$, i.e., the cost of C_1 is doubled.

We first observe that, as one would expect, the trajectories of the optimal capacity vectors over the entire time horizon follow the demand pattern of the traffic intensity vectors; see Figure 2 and the leftmost plots of Figures 3 – 5. Some of the interactions among different optimal capacity solutions are investigated through a comparison of the effects of doubling the cost for class 1 capacity. In comparing the leftmost and center plots of Figures 3 – 5, we observe that while doubling the cost K_1 of link 1 may affect the optimal capacities of link 1 to the greatest extent, it also impacts the optimal capacities of links 2 and 3 over the time horizon, and it does so in different ways under the different demand profiles. These results indicate the delicate interplay among the effects of capacity allocation to various links in the stochastic loss network.

The interactions among different optimal capacity vectors are further investigated through a comparison of the effects of changing the mean call duration lengths. For the case of unequal durations in the rightmost plots of Figures 3 – 5, we observe that since the mean call duration length for route r_1 is twice that of route r_2, the expected loss probability L_1 for calls on this route will be greater and, as such, revenues from calls on route r_1 can be intuitively expected to be lower than those from calls on route r_2. Hence, we would expect the optimal solution to shift more capacity towards link 1 and link 2 which can be seen to use the more remunerative route r_2 from the matrix $\mathbf{A}$. These effects can be indeed observed by comparing the optimal trajectories of C_1 and C_2 in the leftmost and rightmost plots of Figures 3 – 5. We further observe a shift in optimal capacity allocation away from C_3 upon comparing the leftmost and rightmost plots of Figure 4, with more subtle changes exhibited in the leftmost and rightmost plots of Figures 3 and 5.

6. CONCLUSIONS

Motivated by applications from many different areas, we extended the study of the classical stochastic loss network. From the modeling perspective, by allowing the workload process to be modulated as a Markov process with time-varying transition probabilities, we expand the domain of problems that can be modeled by loss networks. From the perspective of solution techniques, we integrate the traditional loss network methodologies, such as the Erlang fixed-point method, with stochastic dynamic programming to conduct a systematic analysis and control of generalized stochastic loss networks. This systematic analysis includes the asymptotic behavior of systems in the limiting regime and the structural properties of very general optimization problem solutions.

The systematic treatment also allows us to further extend our analysis to include additional features that are often encountered and important in applications. For example, quality of serice (QoS) constraints are very commonly adapted in service-related performance analysis and optimization. In the time-varying stochastic loss networks studied in this paper, we can also introduce constraints on the loss rate for each class of customer requests in the profit maximization problem. By the same arguments as those in Section 3, we can show that the solutions to the linear programs with additional constraints is still asymptotically optimal for stochastic systems with the corresponding constraints on the per-class loss rates, and we are in the process of generalizing this result to the general concave utility function version of the optimization problem. Another research topic we are focusing on is to further relax the distributional constraints on the Erlang model, especially in the case of the asymptotic analysis. We also intend to explore the rich theory of the central limit theorems and multivariate Gaussian processes to obtain fixed-point methods under much more general assumptions on the arrival process.

Acknowledgments

The authors thank the referees for helpful comments on an early version of the paper. The authors also thank Sanjay Shakkottai for fruitful discussions and comments on applications of this work in the field of networking.

7. REFERENCES

[1] E. J. Anderson and P. Nash. *Linear Programming in Infinite-Dimensional Spaces: Theory and Applications*. John Wiley and Sons, 1987.

[2] A. Bassamboo, J. M. Harrison, and A. Zeevi. Dynamic routing and admission control in high volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems Theory and Appls.*, 51:249–285, 2006.

[3] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley and Sons, 2nd edition, 1993.

[4] S. L. Bell and R. J. Williams. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *Ann. Appl. Prob.*, 11:608–649, 2001.

[5] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Second edition, 1999.

[6] D. P. Bertsekas. *Dynamic Programming and Optimal Control, Volume II*. Athena Scientific, 2nd edition, 2001.

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

[8] T. Bonald. The Erlang model with non-Poisson call arrivals. In *Proc. Joint SIGMETRICS/Performance Conf. Meas. and Model. Comp. Systems*, pp. 276–286, 2006.

[9] D. Y. Burman, J. P. Lehoczky, and Y. Lim. Insensitivity of blocking probabilities in a circuit-switching network. *J. Appl. Prob.*, 21:850–859, 1984.

[10] G. B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, 1963.

[11] A. K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. In E. Brockmeyer, H. L. Halstrom, and A. Jensen, editors, *The Life and Works of A. K. Erlang*. Academy of Technical Sciences, Denmark, 1948.

[12] L. Green and P. Kolesar. The pointwise stationary approximation for queues with non-stationary arrivals. *Man. Sci.*, 37(2):84–97, 1991.

[13] A. Greenberg, R. Srikant, and W. Whitt. Resource sharing for book-ahead and instantaneous-request calls. In *Proc. ITC 15*, pages 539–548, 1997.

[14] A. A. Jagers and E. A. V. Doorn. On the continued Erlang loss function. *Op. Res. Letters*, 5(1):43–46, 1986.

[15] F. P. Kelly. Blocking probabilities in large circuit-switched networks. *Adv. Appl. Prob.*, 18(2):473–505, 1986.

[16] F. P. Kelly. Routing in circuit-switched networks: Optimization, shadow prices and decentralization. *Adv. Appl. Prob.*, 20(1):112–144, 1988.

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

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

[19] Y. Lu and J. S. Song. Order-based cost optimization in assemble-to-order systems. *Op. Res.*, 53(1):151–169, 2005.

[20] A. A. Puhalskii and M. I. Reiman. A critically loaded multirate link with trunk reservation. *Queueing Systems Theory and Appls.*, 28:157–190, 1998.

[21] S. M. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, 1983.

[22] B. A. Sevastyanov. An ergodic theorem for Markov processes and its application to telephone systems with refusals. *Theoretical Prob. Appls.*, 2:104–112, 1957.

[23] W. Whitt. Blocking when service is required from several facilities simultaneously. *AT&T Bell Laboratories Technical Journal*, 64(8):1807–1856, 1985.

[24] W. Whitt. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rate increases. *Man. Sci.*, 37(2):307–314, 1991.

[25] D. Wischik and A. Greenberg. Admission control for booking ahead shared resources. In *Proc. IEEE INFOCOM 98*, volume 2, pages 873–882, 1998.

[26] S. Xu, J. S. Song, and B. Liu. Order fulfillment performance measures in an assemble-to-order system with stochastic leadtime. *Op. Res.*, 47(1):131–149, 1999.

[27] G. Yin and Q. Zhang. *Discrete-Time Markov Chains: Two-Time-Scale Methods and Applications*. Springer-Verlag, 2005.

[28] P. H. Zipkin. *Foundations of Inventory Management*. McGraw-Hill, 2000.

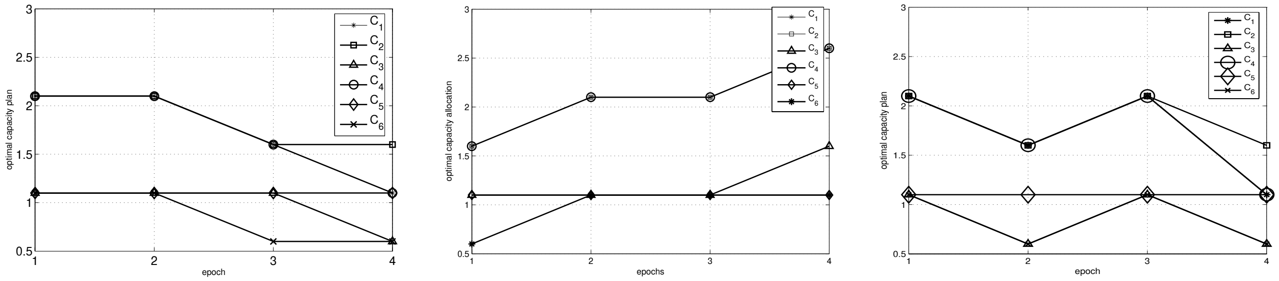

Figure 2: Optimal Capacities under demand Profiles 1–3 on Network 2.

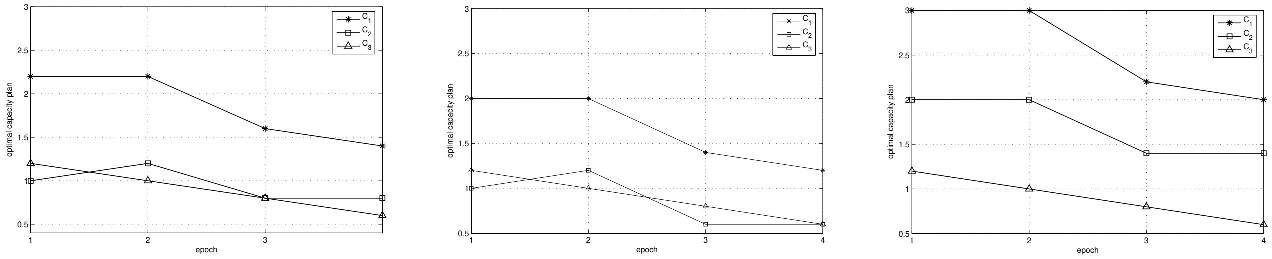

Figure 3: Optimal Capacities under demand Profile 1 on Network 1, for the base case (left), $K_1' = 2K_1$ (center) and $\mu_2' = 2\mu_2$ (right), respectively.

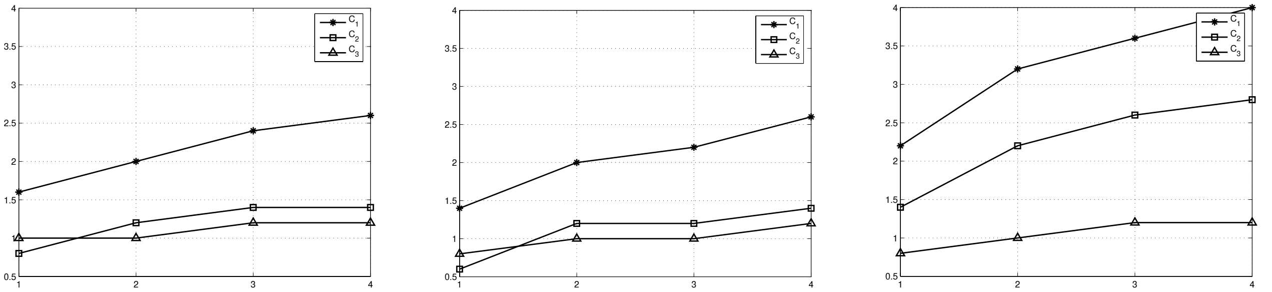

Figure 4: Optimal Capacities under demand Profile 2 on Network 1, for the base case (left), $K_1' = 2K_1$ (center) and $\mu_2' = 2\mu_2$ (right), respectively.

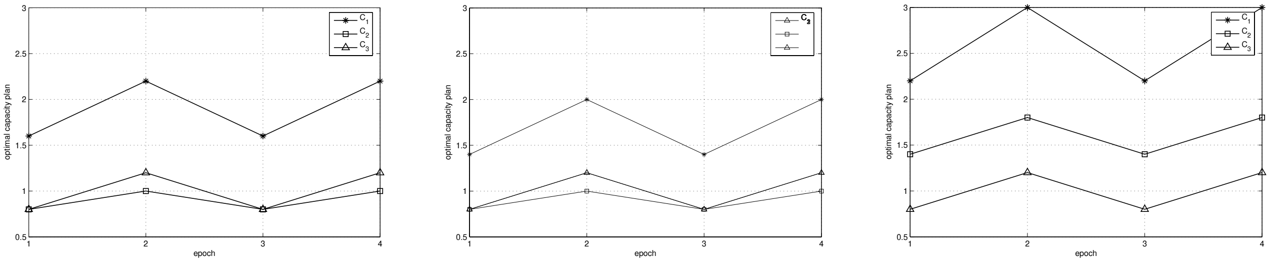

Figure 5: Optimal Capacities under demand Profile 3 on Network 1, for the base case (left), $K_1' = 2K_1$ (center) and $\mu_2' = 2\mu_2$ (right), respectively.