RC25018 (W1006-101) June 23, 2010
Mathematics

IBM Research Report

Stochastic Analysis and Optimization of Multiserver Systems

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

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

Stochastic Analysis and Optimization of Multiserver Systems

Mark S. Squillante

Abstract. Motivated by emerging trends and applications such as autonomic computing, this paper presents an overview of some research in the stochastic analysis and optimization of multiserver systems. Our primary focus is on multiserver systems in general, since this research provides the mathematical methods and results that have been and will continue to be used for the stochastic analysis and/or optimization of existing and future multiserver systems arising in a wide variety of application domains including autonomic computing.

Mathematics Subject Classification (2000). Primary 60G20, 65K10, 93E03; Secondary 60K25, 68M20, 90B15, 90B36, 90C30.

Keywords. Stochastic analysis, stochastic optimization/control, multiserver systems, multidimensional stochastic processes.

1. Introduction

In the beginning there was the single-server queue. And the queue was in its simplest form, and void of known results. And man studied the single-server queue to let there be light upon this darkness. And man derived its mathematical properties and applied these results to the design, analysis and optimization of computer systems and networks. And the derivations of these mathematical results and their applications proved to be fruitful and they multiplied. And the book of Cohen [16]*, from the family of Temple priests (kohanim), was the authoritative text on this subject, often referred to as the bible of the single-server queue, presenting and deriving some of the most fundamental results in the area. And man saw that the single-server queue was good, both in theory and in practice.*

Those who know the author will not require any explanation that the above analogy is intended to make a serious point, and those who do not know the author should understand that no disrespect is intended in *any* way. Our intention here is to highlight the important role that the single-server queue often played in the genesis of the stochastic analysis and optimization of multiserver systems in practice. As a specific example, from the earliest days of computing up until the last decade or so, there was a continual

debate among computer architects about whether improved performance in computer designs should be achieved through increasing the speed of a single centralized processor or through the use of multiple processors, with the decision always being made in favor of the single-server design approach (with the exception, of course, that multiserver computer systems were indeed built, but they were in the vast minority and were built for other reasons) [76]. Many of the most important reasons for this consistent design choice were based (consciously or not) on the mathematical properties and optimization results obtained for the single-server and multiserver queueing systems under the type of scheduling policies (relatively simple timesharing) and workloads (not involving heavy-tailed service time distributions) found in the computer systems of the day. The more recent switch to multiserver computer designs by computer architects over the past decade or so, with multiple processors on each of the multiple chips comprising the computer, has been the result of constraints due to physics and power consumption and changes in the objective function rather than the fundamental properties established for single-server and multiserver systems [76].

On the other hand, the interest in and development of multiserver systems has moved far beyond its initial role as a natural alternative design to single-server systems. New and emerging trends in technology and a wide variety of applications have created a significant increase in both the level and breadth of interest in the stochastic analysis and optimization of multiserver systems. One particularly important recent and emerging trend in technology and applications, as well as the focus of the present book, is autonomic computing. An autonomic computing system is a complex computing environment comprised of many interconnected components that operate at different time scales in a largely independent fashion and that manage themselves to satisfy high-level system management and performance requirements. Autonomic computing systems are also dynamic environments in which optimal self-management decisions must be made continually over time and at multiple time scales. Fundamental problems involved in achieving the goals of autonomic computing concern a general mathematical framework that provides the underlying foundation and supports the design, architecture and algorithms of the decision making components employed throughout the autonomic computing system. At the most basic level, such autonomic computing environments are general multiserver systems of various forms which reflect the increasing complexity of current and future computing systems. Hence, a fundamental aspect of the desired mathematical framework is the stochastic analysis and optimization of multiserver systems in general, where autonomic computing as well as other emerging trends have created a significant increase in the complexity and diversity of the multiserver systems of interest with respect to the analysis and optimization of such stochastic systems.

We therefore consider in this paper some fundamental approaches, methods and results comprising a mathematical framework for the general stochastic analysis and optimization of multiserver systems over time, including the complexities and difficulties at various time scales of autonomic computing and other emerging trends in technology. While stochastic analysis and optimization each play a predominant role, it can be difficult some times to separate these aspects of the desired mathematical framework to a great extent. In many cases, the stochastic optimization of a multiserver system can be based on

a stochastic analysis of the multiserver system over which the optimization is performed. Similarly, once one derives a stochastic analysis of a multiserver system, it can be quite natural to then want to perform a stochastic optimization of the multiserver system upon gaining insights through this analysis.

The overwhelming breadth and depth of the relevant research literature on the stochastic analysis and optimization of multiserver systems prohibits an exhaustive exposition, and thus we do not even attempt to do so. We do attempt to consider a broad range of approaches, methods and results that have been and will continue to be used in the stochastic analysis and optimization of existing and future multiserver systems, as motivated by autonomic computing and other emerging applications. However, this paper considers only a very small fraction of the relevant research on the stochastic analysis and optimization of multiserver systems. We focus on explicit mathematical models, and in particular stochastic models, of general multiserver systems. Even within this context, a number of important areas are not covered at all, such as the vast research on many server systems motivated by call centers and other service operations management systems; see, e.g., [35]. Once again, the subject matter is simply far too broad and deep for us to provide an exhaustive exposition. Finally, we refer the interested reader to two very nice survey papers [12, 1] and the references therein for additional research studies related to the stochastic analysis and optimization of multiserver systems.

The paper is organized as follows. We first summarize the general multiserver model and some mathematical definitions and results used in the paper. Instead of being spread throughout the paper, we centralize this material in Section 2 for easier reference. The next two sections primarily consider exact methods and results, where Sections 3 and 4 focus on boundary value problems and stability, respectively. Section 5 considers both exact and approximate approaches, whereas approximations based on limiting regimes are considered in Section 6. A few issues related to decentralized control and dynamics are briefly discussed in Section 7, followed by some concluding remarks.

2. Technical Preliminaries

2.1. Generic Model Description

We consider a generic multiserver system consisting of S servers in which customers arrive according to an exogenous stochastic process $A(t)$ with mean interarrival time $\lambda^{-1} = \mathbb{E}[A]$ and customer service times on server $s = 1, \ldots, S$ follow a stochastic process $B_s(t)$ with mean $\mu_s^{-1} = \mathbb{E}[B_s]$. In multiclass instances of this generic multiserver system, customers of class $c = 1, \ldots, C$ arrive according to an exogenous stochastic arrival process $A_c(t)$ with mean interarrival time $\lambda_c^{-1} = \mathbb{E}[A_c]$ and class c customer service times on server $s = 1, \ldots, S$ follow a stochastic process $B_{sc}(t)$ with mean $\mu_{sc}^{-1} = \mathbb{E}[B_{sc}]$. We allow $\mathbb{E}[A] = \infty$ and $\mathbb{E}[A_c] = \infty$, in which case the corresponding exogenous arrival process is not considered, and we allow $\mathbb{E}[B_s] = \infty$ and $\mathbb{E}[B_{sc}] = \infty$, in which case the corresponding service process is not considered. Let $Q_i(t)$ denote the number of type-i customers in the multiserver system at time t, and let $\mathbf{Q}(t) = (Q_i(t))_{i \in \mathcal{Q}}$ be the corresponding number in system vector (often the queue length vector process), where

the index i can represent a server or customer class or combination of both with the set of such indices denoted by $\mathcal{Q}$. Define $\mathbf{Q} = \{\mathbf{Q}(t)\,;\, t \geq 0\}$ to be the corresponding multidimensional stochastic number in system process for the multiserver system. Further assumptions can be, and typically are, imposed on the above stochastic processes, but we instead focus on a generic multiserver system and consider the stochastic analysis and optimization of these systems in general, leaving it to the references to provide the additional assumptions associated with any specific results.

A wide variety of structural organizations and topologies exist for such generic multiserver systems and this continues to grow. These organizations and topologies include a single queue of customers being served by a set of servers, through a single-tier of multiple servers that service multiple queues of different classes of customers, up to a network of single-server queues or multiserver queues in either of these forms under arbitrary organizations and topologies, as well as every possibility in between and any possible combination. The servers can be homogeneous or heterogeneous. Upon completing the service of a customer, the server follows a scheduling policy to determine which customer to serve next, including the possibility of remaining idle even when customers are waiting as the policy need not be work conserving. Upon completing its service at a server, the customer follows a routing policy to determine whether it leaves the system or moves to one of the system queues to receive service, possibly switching to another customer class. Once again, we make no specific assumptions about the scheduling or routing policies employed in the multiserver system, leaving it to the references to provide additional assumptions associated with any specific results. Our interests in this paper span the entire spectrum of multiserver systems in general and most of the statements in the paper will correspond to this entire spectrum of multiserver organizations and topologies. Any statements intended for a specific organization or topology should be clear from the context.

2.2. Mathematical Definitions and Results

In this section we briefly summarize some mathematical definitions and results used throughout the paper. These mathematical methods and results can play an important role in the analysis and optimization of multiserver systems, as we shall see in subsequent sections; they can equally play an important role in the analysis and optimization of multidimensional stochastic models in general. Many technical details are omitted and we refer the interested reader to the references provided. Let $\mathbb{R}^+$ ($\mathbb{R}_+$) and $\mathbb{Z}^+$ ($\mathbb{Z}_+$) denote the set of positive (nonnegative) reals and integers, respectively, and define $\mathbf{e}$ to be a column vector of proper order containing all ones.

Consider a discrete-time Markov process $\mathbf{X}^{\mathrm{o}} = \{\mathbf{X}^{\mathrm{o}}(t)\,;\, t \in \mathbb{Z}_+\}$ on a countable, multidimensional state space $\mathcal{X}$. The definition of a corresponding Lyapunov function can be stated as follows. A nonnegative function $\Phi : \mathcal{X} \to \mathbb{R}_+$ is a *Lyapunov function* if there exist some $\gamma > 0$ and $B \geq 0$ such that for any $t \in \mathbb{Z}^+$ and any $\mathbf{x} \in \mathcal{X}$, with $\Phi(\mathbf{x}) > B$,

$$\mathbb{E}[\,\Phi(\mathbf{X}^{\mathrm{o}}(t+1)) \,|\, \mathbf{X}^{\mathrm{o}}(t) \,=\, \mathbf{x}\,] \;\leq\; \Phi(\mathbf{x}) \;-\; \gamma.$$

Refer to, e.g., [62, 29] for additional technical details.

Consider a discrete-time Markov process $\mathbf{X}^{\mathrm{o}} = \{\mathbf{X}^{\mathrm{o}}(t)\,;\, t \in \mathbb{Z}_+\}$ on a countable, multidimensional state space $\mathcal{X} = \bigcup_{i=1}^{L} \mathcal{X}_i$ with transition probability matrix $\mathbf{T}^{\mathrm{o}}$ having

the form

$$\mathbf{T}^{\mathrm{o}} = \begin{pmatrix} \mathbf{P}^{\mathrm{o}}_{11} & \mathbf{P}^{\mathrm{o}}_{12} & \cdots & \mathbf{P}^{\mathrm{o}}_{1L} \\ \mathbf{P}^{\mathrm{o}}_{21} & \mathbf{P}^{\mathrm{o}}_{22} & \cdots & \mathbf{P}^{\mathrm{o}}_{2L} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{P}^{\mathrm{o}}_{L1} & \mathbf{P}^{\mathrm{o}}_{L2} & \cdots & \mathbf{P}^{\mathrm{o}}_{LL} \end{pmatrix}, \tag{2.1}$$

where $\mathbf{P}^{\mathrm{o}}_{ik}$ has dimension $|\mathcal{X}_i| \times |\mathcal{X}_k|$, $i, k = 1, \ldots, L$. The matrix $\mathbf{P}^{\mathrm{o}}_{ik}$ defines the transitions from states in $\mathcal{X}_i$ to states in $\mathcal{X}_k$, $i, k = 1, \ldots, L$, and L denotes the number of block partitions of $\mathbf{T}^{\mathrm{o}}$. Define for $\mathbf{x}_{i,j} \in \mathcal{X}_i$, $j \in \{1, \ldots, |\mathcal{X}_i|\}$, $i = 1, \ldots, L$,

$$\begin{aligned} \pi(\mathbf{x}_{i,j}) &\triangleq \lim_{t\to\infty} \mathbb{P}[\mathbf{X}^{\mathrm{o}}(t) = \mathbf{x}_{i,j}], \\ \boldsymbol{\pi}_i &\triangleq (\pi(\mathbf{x}_{i,1}), \pi(\mathbf{x}_{i,2}), \ldots, \pi(\mathbf{x}_{i,|\mathcal{X}_i|})), \\ \boldsymbol{\pi} &\triangleq (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_L). \end{aligned}$$

The limiting probability vector $\boldsymbol{\pi}$ is the stationary distribution of the stochastic process $\mathbf{X}^{\mathrm{o}}$, which we assume to be irreducible and ergodic and thus the stationary distribution is uniquely determined by solving the global balance equations $\boldsymbol{\pi}\mathbf{T}^{\mathrm{o}} = \boldsymbol{\pi}$ and the normalizing constraint $\boldsymbol{\pi}\mathbf{e} = 1$.

Consider a continuous-time Markov process $\mathbf{X} = \{\mathbf{X}(t)\,;\, t \in \mathbb{R}_+\}$, on a countable, multidimensional state space $\mathcal{X} = \bigcup_{i=0}^{\infty} \mathcal{X}_i$ with infinitesimal generator matrix $\mathbf{T}$ having the form

$$\mathbf{T} = \begin{pmatrix} \mathbf{B}_{00} & \mathbf{B}_{01} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots \\ \mathbf{B}_{10} & \mathbf{B}_{11} & \mathbf{A}_0 & \mathbf{0} & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{2.2}$$

where $\mathbf{B}_{00}$, $\mathbf{B}_{01}$, $\mathbf{B}_{10}$, $\mathbf{A}_{k:k=0,1,2}$, have dimensions $|\mathcal{X}_B| \times |\mathcal{X}_B|$, $|\mathcal{X}_B| \times |\mathcal{X}_N|$, $|\mathcal{X}_N| \times |\mathcal{X}_B|$, $|\mathcal{X}_N| \times |\mathcal{X}_N|$, respectively, with $\mathcal{X}_B = \bigcup_{i=0}^{N-1} \mathcal{X}_i$. The matrix $\mathbf{A}_2$ defines the transitions from states in $\mathcal{X}_i$ to states in $\mathcal{X}_{i-1}$, $i \in \{N+1, N+2, \ldots\}$, $\mathbf{A}_0$ defines the transitions from states in $\mathcal{X}_i$ to states in $\mathcal{X}_{i+1}$, $i \in \{N, N+1, \ldots\}$, and the off-diagonal elements of $\mathbf{A}_1$ define the transitions between states within $\mathcal{X}_i$, $i \in \{N+1, N+2, \ldots\}$. Define for $\mathbf{x}_{i,j} \in \mathcal{X}_i$, $j \in \{1, \ldots, |\mathcal{X}_i|\}$, $i \in \mathbb{Z}_+$,

$$\begin{aligned} \pi(\mathbf{x}_{i,j}) &\triangleq \lim_{t\to\infty} \mathbb{P}[\mathbf{X}(t) = \mathbf{x}_{i,j}], \\ \boldsymbol{\pi}_i &\triangleq (\pi(\mathbf{x}_{i,1}), \pi(\mathbf{x}_{i,2}), \ldots, \pi(\mathbf{x}_{i,|\mathcal{X}_i|})), \\ \boldsymbol{\pi} &\triangleq (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots). \end{aligned}$$

The limiting probability vector $\boldsymbol{\pi}$ is the stationary distribution of the stochastic process $\mathbf{X}$, which we assume to be irreducible and ergodic and thus the stationary distribution is uniquely determined by solving the global balance equations $\boldsymbol{\pi}\mathbf{T} = \mathbf{0}$ and the normalizing constraint $\boldsymbol{\pi}\mathbf{e} = 1$. From standard matrix-analytic analysis, the stationary distribution

$\boldsymbol{\pi}$ has a matrix-geometric form given by

$$\boldsymbol{\pi}_{N+n} = \boldsymbol{\pi}_N \mathbf{R}^n, \qquad n \in \mathbb{Z}_+, \tag{2.3}$$

$$\mathbf{0} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N) \begin{pmatrix} \mathbf{B}_{00} & \mathbf{B}_{01} \\ \mathbf{B}_{10} & \mathbf{B}_{11} + \mathbf{R}\mathbf{A}_2 \end{pmatrix}, \tag{2.4}$$

$$1 = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{N-1})\, \mathbf{e} + \boldsymbol{\pi}_N (\mathbf{I} - \mathbf{R})^{-1} \mathbf{e}, \tag{2.5}$$

where $\mathbf{R}$ is the minimal nonnegative matrix that satisfies $\mathbf{R}^2\mathbf{A}_2 + \mathbf{R}\mathbf{A}_1 + \mathbf{A}_0 = \mathbf{0}$. Refer to, e.g., [67, 68, 56] for additional details.

Consider a continuous-time Markov process $\mathbf{X} = \{\mathbf{X}(t) \,;\, t \in \mathbb{R}_+\}$, on a countable, multidimensional state space $\mathcal{X}$. The *fluid limit* of this process is associated with the almost sure convergence of the scaled process $\widetilde{\mathbf{X}}^n(t) = \mathbf{X}^n(nt)/n$ as $n \to \infty$ such that

$$\widetilde{\mathbf{X}}^n \;\to\; \widetilde{\mathbf{X}}, \qquad \text{u.o.c.,} \qquad\qquad \text{as } \; n \to \infty.$$

Similarly, the *diffusion limit* of the stochastic process $\mathbf{X}$ is associated with the weak convergence of the scaled process $\widehat{\mathbf{X}}^n(t) = \mathbf{X}^n(nt)/\sqrt{n}$ as $n \to \infty$ such that

$$\widehat{\mathbf{X}}^n \;\stackrel{d}{\to}\; \widehat{\mathbf{X}}, \qquad\qquad \text{as } \; n \to \infty.$$

Refer to, e.g., [15, 93] for additional technical details.

Consider a sequence of multiserver systems, indexed by $n = 1, 2, \ldots$, where the nth system operates under the control policy $\mathbb{K}^n$, in the heavy traffic limit (commensurate with diffusion scaling of the associated underlying stochastic processes) as $n \to \infty$. Let $J^n(\mathbb{K}^n)$ be the expected cost for the nth multiserver system under the control policy $\mathbb{K}^n$. Then a control policy $\mathbb{K}^{n,*}$ is called *asymptotically optimal* if for any feasible policy $\mathbb{K}^n$, we have

$$\liminf \frac{J^n(\mathbb{K}^n)}{J^n(\mathbb{K}^{n,*})} \;\geq\; 1, \qquad \text{as } n \to \infty. \tag{2.6}$$

This definition indicates that the cost $J^* = \lim_{n\to\infty} J^n(\mathbb{K}^{n,*})$ is the best cost one can achieve asymptotically and that this asymptotically minimal cost is achieved by the sequence of control policies $\{\mathbb{K}^{n,*}\}$. Refer to, e.g., [7] for additional technical details.

3. Boundary Value Problems

The stochastic analysis and optimization of multiserver systems often involve the analysis of Markov processes defined on countable, multidimensional state spaces. This general class of multidimensional problems is notoriously difficult to solve exactly with analytic solution methods. In fact, these multidimensional aspects of the stochastic process underlying the multiserver system is one of the major sources of complexity and difficulty in the stochastic analysis and optimization of multiserver systems. On the other hand, a number of general approaches have been developed to solve certain instances of two-dimensional multiserver systems.

In one well-known example of the so-called two coupled processor model [27], it has been shown that the functional equations for the two-dimensional generating function of the joint queue length distribution can be reduced to a Riemann-Hilbert boundary value

problem, making it possible to exploit results from the general theory of boundary value equations and singular integral equations. Systematic and detailed studies of this general approach and its use in the stochastic analysis and optimization of distinct multiserver systems can be found in [19, 28]. Some additional applications of this approach include shortest queue routing, fork-join queues, the so-called 2×2 switch, two-dimensional random walks, and the M/G/2 queue. We refer the interested reader to [19, 18, 28, 1] and the references cited therein, noting that other related general approaches are discussed in [1].

Another application of this approach is the classical longest queue model in which a single server always serves the longest of two queues with ties broken in a probabilistic manner. The relevant functional equation for a version of the longest queue model is reduced to a Riemann boundary value problem in [17], which also includes a derivation of the solution of this boundary value problem. Another version of the longest queue model is considered in [30] and [95], where the former determines the limiting probabilities for a corresponding Markov process by solving a functional equation for the generating function obtained from the relevant balance equations and the latter determines these limiting probabilities directly from the balance equations. More recently, an explicit solution for the stationary distribution of the longest queue model has been obtained in [90] based on a matrix-analytic analysis, in terms of versions of (2.3) – (2.5), where explicit expressions for the elements of the $\mathbf{R}$ matrix are determined through the solution of a corresponding lattice path counting problem derived using path decomposition, Bernoulli excursions and hypergeometric functions. The results in [90] also support the multi-server version of this longest queue model, and further provide explicit solutions for the stationary distribution of a general class of random walks in the quarter-plane (namely, $\mathbb{Z}_+^2$ [28]).

4. Stability and Throughput

The stability of multiserver systems represents important issues in the stochastic analysis and optimization of such systems. Stability is also directly related to the maximum throughput of multiserver systems, which is often an important performance objective for the design and optimization of these multiserver systems. Moreover, the rate at which the maximum throughput of a multiserver system scales with respect to the number of servers S as $S \to \infty$ is another important topic of both theoretical and practical interest for the analysis and optimization of multiserver systems.

The stability of multiserver systems has been a fundamental aspect of the stochastic analysis of these systems from the very beginning, with the stability conditions also providing the maximum throughput of the system. In recent years, the issue of stability has received a great deal of attention, especially with respect to single-class and multiclass queueing networks. This recent interest was piqued by several studies showing that the traditional stability condition, namely that the nominal load at each queue/server is less than unity, is not sufficient for a large class of multiserver systems under various scheduling policies. (For example, mutual blocking among the servers can cause such instabilities; see, e.g., [53, 13] and the references cited therein.) A wide variety of methods

and results have been developed to address the stability of multiserver systems, and we refer the interested reader to [62, 29] and the references therein for a thorough treatment of much of this research. Of particular interest is the unified approach via fluid limits developed in [22], generalizing the related earlier work in [78], based on the key result that a queueing network is stable if the corresponding fluid limit network is stable in the sense that the fluid network eventually reaches zero and stays there regardless of the initial multiserver system configuration. This approach and related extensions have played an important role in determining the stability conditions of multiserver systems, and the design of optimal scheduling policies, especially since the analysis can focus on the fluid limit of the multiserver system rather than the more complex stochastic system.

Due to the explosive growth in wireless technology and applications, the asymptotic rate at which the maximum throughput of wireless networks scales with respect to the size of the network S has become an important theoretical and practical issue. A random multiserver model of static wireless networks was used in [37] to show that the maximum throughput per source-destination pair is $O(1/\sqrt{S})$ as $S \to \infty$. Also presented is a $\Theta(1/\sqrt{S \log S})$ throughput scheme, which has been generalized to a parametrized version that achieves the optimal throughput-delay tradeoff for maximum throughputs of $O(1/\sqrt{S \log S})$ [24, 25, 26]. See [48, 54, 57] for further extensions of the original model and their analyses. The focus has recently turned to the asymptotic scalability of wireless networks under constant-size buffers at each server of the multiserver system, for which it has been shown that there is no end-to-end protocol capable of achieving the maximum throughput of $O(1/\sqrt{S})$ as $S \to \infty$ [46, 47]. However, it is also shown that there exists a protocol which achieves the asymptotic maximum throughput of $O(1/\sqrt{S \log S})$ with constant-size per-server buffers and which has to employ a local buffer coordination scheduling scheme.

The methods and results used to determine the stability conditions, and in turn maximum throughput, of multiserver systems have also been extended to obtain a broader set of performance metrics through important connections between the stability and the stationary distribution $\boldsymbol{\pi}$ of multiserver systems. As a specific example, a general methodology is proposed in [9] based on Lyapunov functions to study the stationary distribution of infinite multidimensional Markov processes $\mathbf{Q}$, which model a general set of multiclass multiserver systems. This methodology is based on key results showing that if there exist linear or piecewise linear Lyapunov functions which establish the stability of multiserver systems, then these Lyapunov functions can also be used to determine upper and lower bounds on the stationary tail distribution, which in turn provide bounds on the expected queue lengths. These upper and lower bounds hold uniformly under any work conserving policy, and the lower bounds are further extended to priority policies. The results in [9] also represent the first explicit geometric upper and lower bounds on the tail probabilities of the multidimensional queue length process $\mathbf{Q}$ for such general multiserver systems.

In another example related to infinite multidimensional Markov processes [34], more specifically a stochastic online version of the classical bin packing scheduling problem, a stochastic analysis of the corresponding multiserver system is developed based on

a combination of a Lyapunov function technique and matrix-analytic methods. These results include the stability conditions and the stationary distribution $\boldsymbol{\pi}$ of the joint queue length process $\mathbf{Q}$ for general stochastic multidimensional bin packing processes. The stability and stationary distribution results are both derived in a recursive manner by exploiting a priority structural property, where the stability condition for the current level of the partitioned queue length process is obtained using a Lyapunov function technique involving the stationary distribution for the previous level of the partitioned queue length process, and the stationary distribution for the current level is obtained from (discrete-time) versions of (2.3) – (2.5). In addition, various performance metrics are obtained including asymptotic decay rates and expected wasted space, and large deviations bounds are used to obtain an accurate level of truncation. The approach in [34] is also based on a form of stochastic decomposition, which is generally considered in more detail in the next section.

5. Stochastic Decomposition

The multidimensional aspects of stochastic processes underlying multiserver systems are one of the many sources of complexity in their stochastic analysis and optimization, which often involve various dependencies and dynamic interactions among the different dimensions of the multidimensional process. Hence, a considerable number of general approaches have been developed that essentially decompose the complex multidimensional stochastic process into a combination of various forms of simpler processes with reduced dimensionality.

One general class of stochastic decomposition approaches is based on the theory of nearly completely decomposable stochastic systems. Consider a discrete-time Markov process with transition probability matrix $\mathbf{T}^{\text{o}} = [t^{\text{o}}_{\mathbf{x}_{i,j}\,\mathbf{x}_{k,\ell}}]$ in the form of (2.1) and with state space $\mathcal{X}$. When $|\mathcal{X}|$ is very large, computing the stationary distribution (as well as functions of the stationary distribution) directly from the transition probability matrix can be prohibitively expensive in both time and space. Suppose, however, that the block submatrices along the main diagonal (i.e., $\mathbf{P}^{\text{o}}_{11}, \cdots, \mathbf{P}^{\text{o}}_{LL}$) consist of relatively large probability mass, while the elements of the other block submatrices are very small in comparison (i.e., $\mathbf{P}^{\text{o}}_{ik} \approx \mathbf{0}$, $i \neq k$). Matrices of this type are called nearly completely decomposable [20], in which case the matrix $\mathbf{T}^{\text{o}}$ can be written in the form $\mathbf{T}^{\text{o}} = \mathbf{W}^* + \epsilon\mathbf{D}$ where $\mathbf{W}^* = \text{diag}(\mathbf{W}^*_1, \mathbf{W}^*_2, \cdots, \mathbf{W}^*_L)$, the matrices $\mathbf{W}^*_i$ are stochastic and completely decomposable, $i = 1, \ldots, L$, ϵ is small compared to the elements of $\mathbf{W}^*$, and the absolute value of each element of $\mathbf{D}$ is less than or equal to 1. The model solution is then based on extensions of the Simon-Ando approximations for the stationary distribution of the corresponding Markov process. More specifically, given a function $F(\mathbf{T}^{\text{o}})$ of interest, which can include its stationary distribution $\boldsymbol{\pi}$, it follows from the theory of nearly completely decomposable matrices that the function can be approximated as $F(\mathbf{T}^{\text{o}}) \approx \sum_{i=1}^{L} \widetilde{\boldsymbol{\pi}}_i F(\mathbf{W}^*_i)$, the accuracy of which is known to be within $\text{O}(\epsilon)$ [20]. Here, $F(\mathbf{W}^*_i)$ is determined from the matrix $\mathbf{W}^*_i$ and its invariant probabililty vector $\widehat{\boldsymbol{\pi}}_i$, while $\widetilde{\boldsymbol{\pi}}_i$ is determined as the invariant probability vector of the stochastic matrix of dimension $L \times L$ whose elements are given by $\hat{t}^{\text{o}}_{ik} = \sum_{j=1}^{|\mathcal{X}_i|} \sum_{\ell=1}^{|\mathcal{X}_k|} \widehat{\boldsymbol{\pi}}_{ij} t^{\text{o}}_{\mathbf{x}_{i,j}\,\mathbf{x}_{k,\ell}}$. Note that

$\hat{t}^{\rm o}_{ik} = \mathbb{P}[\mathbf{X}^{\rm o}(t+1) \in \mathcal{X}_k | \mathbf{X}^{\rm o}(t) \in \mathcal{X}_i]$, $i,k = 1,\ldots,L$. Error bounds also can be obtained within this framework; see, e.g., [21, 88]. As a specific example, refer to [4] for model instances where $F(\mathbf{T}^{\rm o})$ represents the stationary page fault probability for a computer program model $\mathbf{T}^{\rm o}$ and a finite storage capacity. A solution for instances of these computer storage models with $F(\mathbf{W}^*_i) = 0$, $i = 1,\ldots,L$, while $F(\mathbf{T}^{\rm o}) \neq 0$ is derived in [74, 83] based on first passage times, recurrence times, taboo probabilities and first entrance methods, whereas standard nearly completely decomposable models and analyses obviously break down and fail in such model instances. For additional details on nearly completely decomposable stochastic systems and their solutions, we refer the interested reader to [20] and the references therein.

Another general class of stochastic decomposition approaches is based on models of each dimension of the multidimensional process in isolation together with a fixed-point equation to capture the dependencies and dynamic interactions among the multiple dimensions. In order to consider what is probably the most well-known example of this general approach, let us first recall that the classical Erlang loss model consists of J links, with each link j having capacity C_j, and a set of routes $\mathcal{R}$ defined as a collection of links. Calls for route r arrive with rate λ_r and require capacity $\mathbf{A}_{jr}$ from link j, $\mathbf{A}_{jr} \in \mathbb{Z}_+$. Such a call arrival is lost if the available capacity on any link j is less than $\mathbf{A}_{jr}$, $\forall j = 1,\ldots,J$, and otherwise the call reserves the available capacity $\mathbf{A}_{jr}$ on each link j for a duration having mean μ_r^{-1}, $\forall j = 1,\ldots,J$. The traffic intensity for route r is denoted by $\rho_r = \lambda_r/\mu_r$. It is well known that there exists a unique stationary distribution $\boldsymbol{\pi}$ for the number of active calls on all routes r and that $\boldsymbol{\pi}$ has a product-form solution in terms of the traffic intensities ρ_r. Then the stationary probability L_r that a call on route r is lost can be expressed in terms of this stationary distribution. However, the computational complexity of calculating the exact stationary distribution is known to be $\sharp P$ complete [58], thus causing such calculations to be computationally intractable even for moderate values of J and $|\mathcal{R}|$. We refer the interested reader to [52] and the references therein for additional details.

The well-known Erlang fixed-point approximation has been developed to address this computational complexity and it is based on a stochastic decomposition in which the multidimensional Erlang formula is replaced by a system of J nonlinear equations in terms of the one-dimensional Erlang formula. More specifically, the stationary loss probabilities L_r for routes r are given by

$$L_r \;=\; 1 \;-\; \prod_{j=1}^{J} (1 - B_j)^{\mathbf{A}_{jr}},$$

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

$$B_j \;=\; E\left((1 - B_j)^{-1} \sum_{r=1}^{|\mathcal{R}|} \mathbf{A}_{jr} \rho_r \prod_{i=1}^{J} (1 - B_i)^{\mathbf{A}_{ir}},\, C_j \right), \tag{5.1}$$

with

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

being the Erlang formula for the loss probability of an isolated link of capacity C under traffic from an exogenous stream with intensity ρ. Furthermore, it is well-known that there exists a solution $\mathbf{B} \in [0,1]^J$ of the Erlang fixed-point equations (5.1) and that this solution converges to the exact solution of the original Erlang loss model in the limit as the traffic intensity vector $\boldsymbol{\rho}$ and capacity vector $\mathbf{C}$ are increased together in fixed proportion; see [92, 51, 52]. The corresponding capacity planning optimization problem to maximize profit as a function of the loss probabilities L_r and capacities C_j has been considered within this context [92, 52]. The asymptotic exactness of the Erlang fixed-point approximation and optimization based on this approximation, which follows from an instance of the central limit theorem for conditional Poisson random variables, is an important aspect of this general decomposition approach for the stochastic analysis and optimization of complex multiserver systems, though establishing such results is not always possible. Various extensions of the Erlang loss model and Erlang fixed-point approximation are also possible, including recent results to support less restrictive call arrival processes [11] and on optimal capacity planning under time-varying multiclass workloads [10]. More accurate approximations for the Erlang loss model have also been recently developed; refer to [49, 3].

Another example of this general stochastic decomposition approach was developed in [82, 64] to obtain the stationary distribution $\boldsymbol{\pi}$ of a (symmetric) multiserver system in which a scheduling policy assigns customers to the server where they are served most efficiently and in which a threshold-based scheduling policy manages the tradeoff between balancing the workload among the servers and serving the customers in the most efficient manner. A matrix-analytic analysis of the stochastic processes modeling each server in isolation is derived to obtain the corresponding stationary probability vectors in terms of their arrival and departure processes which are modified to reflect the probabilistic behavior of the other servers. These probability vectors are given by versions of (2.3) – (2.5), where explicit solutions for the elements of the matrix $\mathbf{R}$ are obtained in several instances of the multiserver system. Then the modified arrival and departure processes of each server are expressed in terms of the corresponding stationary probability vector, and the final solution of the system of equations is obtained via a fixed-point iteration. This solution can be shown to be asymptotically exact, in terms of the number of servers S, under certain conditions. The results of this study illustrate and quantify the significant performance benefits of the dynamic threshold-based scheduling policy, particularly at moderate to relatively heavy traffic intensities, but also demonstrate the potential for unstable behavior where servers spend most of their time inefficiently serving customers when thresholds are selected inappropriately. The stochastic analysis in [82, 64] can be used to determine the optimal threshold values for the multiserver system as a function of its parameters. Related (non-symmetric) instances of this multiserver system and this dynamic threshold-based scheduling policy have also been considered within the context of diffusion limiting regimes; refer to Section 6.

Yet another general class of stochastic decomposition approaches is based on exploiting various priority structural properties to reduce the dimensionality of the multiserver system in a recursive manner. Although this general approach was originally developed for single-server systems with multiple queues under a priority scheduling discipline (see, e.g., [36, 45]), it has been extended and generalized in many different ways for the stochastic analysis and optimization of multiserver systems. The basic idea consists of a recursive mathematical procedure starting with the two highest priority dimensions of the process that involves: (i) analyzing the probabilistic behavior of the so-called completion-time process, which characterizes the intervals between consecutive points when customers of the lower priority dimension begin service within a busy period; (ii) obtaining the distributional characteristics of related busy-period processes through an analysis of associated stochastic processes and modified service time distributions in isolation; and (iii) determining the solution of the two-dimensional priority process from a combination of these results. These steps are repeated to obtain the solution for the $(c+1)$-dimensional priority process using the results for the c-dimensional priority process, until reaching the final solution for the original multidimensional stochastic process. Refer to, e.g., [36, 45], and the references cited therein.

One example of this general approach for the stochastic analysis of multiserver systems was discussed at the end of Section 4. Several related extensions of this general approach have been developed for the stochastic analysis and optimization of various multiserver systems, e.g., parallel computing systems under a multiclass gang scheduling policy [85], a (single-class) combination of spacesharing and timesharing policies [81], and different (single-class) dynamic coscheduling policies [87, 86]. These approaches generally exploit distinct priority structures in the underlying multidimensional stochastic process together with the probabilistic behavior of dependence structures and dynamics resulting from the multiserver workloads and policies. More specifically, these approaches investigate each dimension of the stochastic process in isolation based on an analysis of the probabilistic behavior of a set of stochastic processes analogous to the completion-time process together with an analysis of related busy-period processes and modified service time distributions. In [85], this involves deriving expressions for the conditional distributions of the per-class timeplexing-cycle processes (which characterize the intervals between consecutive quanta for a class) given the queue length vectors in terms of the stationary distributions for the other classes. (In a limiting regime, the exact stationary distribution for the queue length process of each class can be obtained in isolation as an alternating service process with vacations representing periods when other classes receive service.) In [81], this involves deriving a first-passage time analysis of the probabilistic behavior of the departure processes associated with the set of timeplexing-cycle processes (which characterize the intervals between consecutive quanta for a customer) to obtain a set of modified service time distributions that incorporate the effects of timesharing. In [87], this involves deriving an analysis of the probabilistic behavior of a set of overall service processes at each server (characterizing the various states that every parallel application can be in) and expressing this probabilistic behavior in terms of the corresponding stationary distributions for the other servers. A fixed-point iteration is used

in all of these cases to solve the resulting system of equations and obtain the stationary distribution of the corresponding multidimensional stochastic process in the form of (2.3) – (2.5). The probability distributions obtained from each stochastic analysis in isolation are either used directly or replaced with more compact (approximate) forms that are constructed by fitting phase-type distributions to match as many moments (and/or other associated probabilistic functionals) of the original distributions as are of interest using any of the best known methods. In particular, classical busy-period results (refer to, e.g., [66, 68]) can be exploited to obtain a more compact (approximate) form for any busy-period distribution. See also [23].

A similar approach was subsequently taken in [40, 39] for the stochastic analysis of customer assignment with cycle stealing in multiserver systems under a central queue or immediate dispatch. The workload consists of two classes denoted by C_1 and C_2. At any given time, a single server is associated with each class and the cycle stealing mechanism allows the server associated with C_2 to serve customers of C_1. In the immediate dispatch case, the stationary distribution for the C_2 process can be determined in isolation using matrix-analytic methods with the solution given by versions of (2.3) – (2.5); the sojourn time moments can be directly obtained using virtual waiting time analysis for the case of Poisson arrivals. Since the servicing of C_1 customers depends upon the C_2 process, the first three moments of the busy and idle periods of the C_2 process are obtained and used to construct corresponding two-stage Coxian distributions with matching moments. The C_1 process is augmented with the approximate busy and idle period distributions of the C_2 process and analyzed in isolation using matrix-analytic methods to obtain the corresponding stationary distribution in the form of (2.3) – (2.5). This analysis of the multiserver system under immediate dispatch is also extended to the case of multiple C_1 servers. Turning to the analysis for the central queue case, there are some differences in the details of the analysis as one would expect, but the basic approach is quite similar. The stationary distribution for the C_2 process can be determined in isolation using matrix-analytic methods where the first C_2 arrival of a busy period either starts service immediately or must wait for the completion of a C_1 customer already in service; the mean sojourn time can be directly obtained, in the case of Poisson arrivals, using known results for the M/G/1 queue with setup times [89]. A stochastic process is formulated to represent the C_1 process together with the probabilistic behavior of various busy periods associated with C_2, where the first three moments of each of the latter random processes are obtained and used to construct corresponding two-stage Coxian distributions with matching moments. The stationary distribution of this process for C_1 customers can be determined using matrix-analytic methods with the solution given by versions of (2.3) – (2.5). The results of these studies demonstrate that cycle stealing can significantly improve the performance of C_1 customers, while the penalty incurred by C_2 customers is relatively small. Performance improvements are found to be greater for both C_1 and C_2 under a central queue than under immediate dispatch.

This approach subsequently evolved into the so-called method of dimensionality reduction that applies to a class of recursive foreground-background stochastic processes, which includes cycle stealing under immediate dispatch, and a class of generalized foreground-background stochastic processes, which includes cycle stealing under a central queue [70].

Two approximations of dimensionality reduction are also proposed in [70], each attempting to reduce the computational complexity of the recursive use of dimensionality reduction by ignoring dependencies to varying degrees (namely, partial and complete independence assumptions) while maintaining reasonable accuracy. The method of dimensionality reduction has been applied to a number of different multiserver systems, including multiserver systems with multiple priority classes [41], threshold-based policies for reducing switching costs in cycle stealing [71, 72], and threshold-based policies for the so-called Beneficiary-Donor model [73].

6. Stochastic Process Limits

The many sources of complexity and difficulty in the stochastic analysis and optimization of multiserver systems often make an exact analysis intractable for numerous instances of multiserver systems. Hence, a considerable number of general approaches have been developed based on an investigation of the underlying stochastic process and associated control problem in some limiting regime.

The analysis of fluid limits of multiserver systems is one important example of this general approach in which the asymptotic behavior of the underlying stochastic process is typically characterized via a functional strong law of large numbers. As such, the stochastic system is approximated by a deterministic system comprised of dynamic continuous flows of fluid to be drained in a manner analogous to the servicing of discrete customers in the original stochastic system. In addition to the methods and results presented in Section 4, this approach and related extensions have played an important role in the analysis and optimization of multiserver systems. One example is developed in [14, 15] to study the optimal dynamic control and scheduling of multiclass fluid networks. An algorithmic procedure is presented that systematically solves the dynamic scheduling problem by solving a sequence of linear programs. Several important properties of this procedure are established, including an example that a globally optimal solution (namely one rendering optimality of the objective function over every point of time) may not exist, and thus the solution procedure is myopic in this respect. The solution procedure generates within a bounded number of iterations a policy, in the form of dynamic capacity allocation among all fluid classes at each node in the network, that consists of a finite set of linear intervals over the entire time horizon and that is guaranteed to yield a stable fluid network.

In another example associated with the dynamic scheduling of multiclass fluid queueing networks [5], an optimal control approach to the optimization of fluid relaxations of multiclass stochastic networks is developed based on the Pontryagin maximum principle and related theory [75, 79]. The maximum principle is used to derive the exact optimal control policies in the fluid limiting regime for several canonical examples of multiserver systems. A numerical method is proposed, based on the structure of the optimal policy, to compute exact solutions for the fluid network optimal control problem using a discrete approximation that is continually refined until the solution no longer improves. Due to the dimensionality difficulties of this exact approach, an efficient approximate algorithm is also developed to compute the fluid optimal control based on a heuristic that learns

from the exact solution of special cases. Numerical experiments illustrate that a pairwise interaction heuristic yields near-optimal policies. More recently, efficient approximation algorithms have been developed for the class of separated continuous linear programming problems that arise as fluid relaxations of multiclass stochastic networks. For example, in [32], a proposed polynomial-time algorithm is shown to provide a solution that, for given constants $\epsilon > 0$ and $\delta > 0$, drains the fluid network with total cost at most $(1+\epsilon)\mathsf{OPT} + \delta$, where OPT is the minimum cost drainage.

Many optimal control problems in multiserver systems can be studied as Markov decision processes. However, the well known difficulty with this approach for some multiserver systems is the so-called curse of dimensionality. In [59, 60], a form of unification is established between the dynamic programming equations of the Markov decision process of a stochastic network control problem and a related total-cost optimal control problem for the corresponding linear fluid network. This and related results in [59, 60] form the basis of a general framework for constructing control algorithms for multiclass queueing networks, with network sequencing and routing problems considered as special cases. Numerical examples are presented showing close similarity between the optimal policy from the proposed framework and the average-cost optimal policy. In [61], the connections between multidimensional Markov decision processes associated with the optimal control of stochastic networks and the corresponding optimal fluid limit control processes are further studied within the context of the control of stochastic networks using state-dependent safety-stocks. For a few canonical examples, it is shown that the proposed policy is fluid-scale asymptotically optimal and approximately average-cost optimal, leading to a new technique to obtain fluid-scale asymptotic optimality for general networks modeled in discrete time. These results are based on the construction of an approximate solution to the average-cost dynamic programming equations using a perturbation of the value function for an associated fluid model.

The analysis of diffusion limits of multiserver systems is another important example of the general approach of this section in which the asymptotic behavior of the underlying stochastic process is typically characterized via a functional central limit theorem. As such, the stochastic processes underlying the multiserver system are approximated by various Brownian motions that describe the heavy-traffic system behavior. A wide variety of methods and results for this diffusion approximation approach have been developed to address the general stochastic analysis and optimization of multiserver systems, and we refer the interested reader to, e.g., [42, 93] and the references therein. Of particular interest is the well-known Halfin-Whitt regime [38, 93], for which certain heavy-traffic limits have been established as the traffic intensity goes to unity and $S \to \infty$ in an S-server queueing system. This framework and related extensions have played an important role in the stochastic analysis and optimization of multiserver systems, with applications in various areas such as large call center environments where resource (agent) capacity planning and scheduling problems have received considerable attention. In one example associated with the dynamic scheduling of multiclass queueing systems in the Halfin-Whitt heavy-traffic regime [44], the Hamilton-Jacobi-Bellman equation associated with the limiting diffusion control problem is shown to have a smooth solution with an optimal policy having a so-called bang-bang control. Several qualitative insights are also derived

from the stochastic analysis, including a square root rule for the capacity planning of large multiserver systems.

Another general class of diffusion approximation approaches have played an important role in the stochastic analysis and optimization of multiserver systems based on solving the corresponding Brownian control problem. As a representative example having received considerable attention, consider a multiserver system in which each class of customers can be served at any one of a (per-class) subset of the servers, with specific class-server service rates. The Brownian control problem associated with the dynamic scheduling of customers in this multiclass parallel-server system to minimize the cumulative holding costs of customers is studied in [43] where, assuming a so-called complete resource pooling condition, a particular discretization method is proposed to find discrete-review policy solutions. (A symmetric version of the general problem, discussed in Section 5, is studied in [82, 64].) Under the same heavy-traffic complete resource pooling assumption, a candidate for an asymptotically optimal control policy in the form of a dynamic threshold policy is proposed in [94] for the original multiserver system. It is then established that this dynamic threshold scheduling policy is asymptotically optimal in the heavy traffic limit under the complete resource pooling condition and that the limiting cost is the same as the optimal cost in the Brownian control problem [7, 8]. Also, for numerical solutions of such control problems in general, refer to [55].

Another example of this general approach consists of first determining derivatives of the performance function of interest at $\rho = 0$, using a Taylor expansion of the function near $\rho = 0$, then determining the diffusion limit of the underlying stochastic process, and finally obtaining a closed-form approximation for the performance function (e.g., the expected sojourn time in the multiserver system) by interpolating between these light-traffic and heavy-traffic limits. This approach was originally proposed in [77] where the 0^{th} through $n-1^{\text{st}}$ order light-traffic derivatives are combined with the heavy-traffic limit to obtain an n^{th} degree polynomial in ρ as an approximation to the normalized performance function, which in turn is used to produce the desired closed-form approximation. Several instances of this general approach have been developed for the stochastic analysis and optimization of various multiserver systems, including symmetric fork-join queueing systems (see below) [91] and optimal resource allocation in parallel-server systems [84].

Several other multiserver systems have been studied in various limiting regimes. One example is shortest queue routing systems in which each of the S servers has its own dedicated queue and customers join the queue with the shortest length at the instant of their arrival. An analysis of the shortest queue system based on a diffusion limit approximation is presented in [33], whereas an exact analysis for the two-server case [2] and mean sojourn time approximations [63] have also been obtained. A related optimal multiclass scheduling problem is studied in [80] together with the associated sequencing problem at each of the S servers. Another example is fork-join queueing systems in which each server has its own dedicated queue and each customer arrival forks into S tasks, with the ith task assigned to the ith server, such that the customer departs the system only after all of its tasks have received service. An analysis of the fork-join queueing system using an interpolation approximation based on light and heavy traffic limits is presented in [91],

whereas an analysis for the two-server case [31] and bounds on various performance metrics [65, 6] have also been obtained.

7. Decentralized Control and Dynamics

Another important source of complexity and difficulty in the stochastic analysis and optimization of multiserver systems often arises as a result of the decentralized management of (large-scale) environments comprised of a collection of multiserver systems. Hence, an additional number of fundamental issues need to be taken into account in the stochastic analysis and optimization of such multiserver systems over time.

One particularly important issue concerns the quality of a decentralized optimization of the entire collection of multiserver systems in comparison with a globally optimal solution of a centrally managed instance of this entire system; see, e.g., [69]. More specifically, consider a hierarchical system where the first level of the hierarchy consists of n multiserver systems, each of which in turn is the root of a subhierarchy of multiserver systems. A utility function $f_i(x_i, r_i, u_i)$ is associated with the ith multiserver system of the first level, where x_i is the set of variables (including policies) that can be changed or affected in multiserver system i, r_i is the set of resources allocated to multiserver system i, and u_i is the set of external variables (including workloads) that impact multiserver system i. The total utility function for the entire hierarchical system is given by $h(f_1(x_1, r_1, u_1), \ldots, f_n(x_n, r_n, u_n))$, such that h aggregates the utility of each multiserver system of the first level into a single total utility. Then the overall goal of the collection of multiserver systems is to globally optimize the total utility function h among all feasible resource allocations $r_1, \ldots, r_n$ and all feasible sets of variables (policies) $x_1, \ldots, x_n$, yielding total utility h_c. Namely, we have

$$h_c \;=\; \min_{x_i, r_i} h(f_1(x_1, r_1, u_1), \ldots, f_n(x_n, r_n, u_n)).$$

On the other hand, the decentralized optimization of this hierarchy of multiserver systems involves each of the n multiserver systems optimizing its local utility function

$$g_i(r_i, u_i) \;=\; \max_{x_i} f_i(x_i, r_i, u_i)$$

among all feasible sets of variables (policies) x_i given the set of resources r_i allocated by the central manager to multiserver system i. In turn, the central manager optimizes the total utility function $h(g_1(r_1, u_1), \ldots, g_n(r_n, u_n))$ among all feasible resource allocations $r_1, \ldots, r_n$ for the collection of multiserver systems, yielding total utility h_d. Namely, we have

$$h_d \;=\; \min_{r_i} h(g_1(r_1, u_1), \ldots, g_n(r_n, u_n)).$$

Then it can be easily shown [69] that as long as the aggregation function h is order preserving (in the sense that $h(x) \geq h(y)$ whenever $x \geq y$ where $x \geq y$ if $x_k \geq y_k$ for all k and $h : \mathbb{R}^n \to \mathbb{R}^m$), the decentralized optimal solution is as good as the centralized optimal solution, i.e., $h_c = h_d$. The same arguments can be applied recursively at each level of the hierarchy with respect to the decentralized optimization of the entire subcollection of multiserver systems. We refer the interested reader to [69] for additional details.

Another fundamental issue that needs to be addressed in the stochastic analysis and optimization of multiserver systems concerns the dynamics of the system over time and at multiple time scales. Various aspects of each multiserver system (e.g., the external variables u_i) can vary over time, and thus the above decentralized optimization decisions may occur on a periodic basis. The time scales at which these decisions are made at each level of the hierarchical collection of multiserver systems depend upon several factors, including the delays, overheads and constraints involved in making changes to decision variables, the service-level agreements and performance guarantees of each multiserver system, and the properties of the underlying (nonstationary) stochastic processes. In such circumstances, it is well known that even very simple (e.g., linear) models, which are only piecewise continuous or contain a feedback element, may exhibit chaotic behavior (in the sense of difficult to predict and qualitatively very sensitive to initial or control conditions) [50]. As an elementary instance in which adding time delay can produce locally unstable behavior (and hence can produce chaos on the larger scale), consider a linear dynamical system $y_{n+1} = (s-d)y_n + D$ with constant parameters, where the new system state depends only on the closest previous state. This system is stable when $|\xi| \leq 1$ and asymptotically stable when $|\xi| < 1$, where $\xi = s - d$ denotes the eigenvalue of the dynamical system. On the other hand, if the balance of $s - d$ is spread over time, we have a system $y_{n+1} = sy_n - dy_{n-1} + D$ in which now the stability condition is that both solutions of $\xi^2 - s\xi + d$ (the characteristic polynomial of the new system) satisfy $|\xi| \leq 1$. Here the growth rate s corresponds to the rate of growth in the backlog of customers in the multiserver systems and the decay rate d corresponds to the resource allocation in the multiserver systems. In the first dynamical system equation, the growth rate and the decay rate cancel each other within the same time interval, and thus we focus on the net effect which, by assumption, is such that the backlog remains bounded. In the presence of time delay as in the second dynamical system equation, the decay rate (or the resource allocation) corresponds to a different time interval than the growth rate (or the customer backlog), which in some cases produces instabilities. When the dynamical system is near such a fix point and it is globally bounded (by some non-linear dependencies) in such a way that the trajectories return to this fix point, then the instability of the fix point produces very chaotic behavior due to the irregular number of iterates involved in returns to this fix point. Chaos can be controllable in special cases, for example many stochastically stable systems exhibit individual chaotic trajectories, but with very well behaved distributions or moments. The transitions from a deterministic regime, where all trajectories are predictable at all times, to a stochastic regime, where most of the trajectories are predictable over long intervals of time, may go through all kinds of uncontrollable evolutions. It is therefore essential for the stochastic analysis and (decentralized) optimization of multiserver systems to determine the types of possible asymptotic behavior and the stability of such behavior under small perturbations of the system, and to conceive of mechanisms exposing the type of behavior in which the system currently resides. For additional details, we refer the interested reader to [69].

8. Conclusions

The genesis of multiserver systems may have been as straightforward extensions and alternatives to single-server systems, but new and emerging trends such as autonomic computing have been driving a significant growth of interest in multiserver systems. This growth has resulted in new formulations and even greater complexities in the multiserver systems in general from both theoretical and practical perspectives. The stochastic analysis and optimization of multiserver systems must address these complexities and difficulties. This will require extensions of existing solution methods and results, including some of the general approaches considered in this paper, but will also require the development of new solutions methods and the derivation of new results in the stochastic analysis and optimization of multiserver systems.

Acknowledgements

The author especially thanks Danilo Ardagna and Li Zhang for their kind invitation to write this paper. He also thanks Alan Hoffman, Yingdong Lu, Baruch Schieber, Mayank Sharma and Shmuel Winograd for helpful comments on an earlier version of the paper, as well as Dan Prener for fruitful discussions regarding some of the computer architecture points in the introduction.

References

[1] I. J. Adan, O. J. Boxma, and J. Resing. Queueing models with multiple waiting lines. *Queueing Systems Theory and Applications*, 37:65–98, 2001.

[2] I. J. Adan, J. Wessels, and W. Zijm. Analysis of the asymmetric shortest queue problem. *Queueing Systems Theory and Applications*, 8:1–58, 1989.

[3] J. Anselmi, Y. Lu, M. Sharma, and M. S. Squillante. Improved approximations for the Erlang loss model. *Queueing Systems Theory and Applications*, 63:217–239, 2009.

[4] O. I. Aven, E. G. Coffman, Jr., and Y. A. Kogan. *Stochastic Analysis of Computer Storage*. D. Reidel, 1987.

[5] F. Avram, D. Bertsimas, and M. Ricard. Fluid models of sequencing problems in open queueing networks: An optimal control approach. In F. Kelly and R. Williams, editors, *Stochastic Networks*, volume IMA 71, pages 199–234. 1995.

[6] F. Baccelli, A. M. Makowski, and A. Shwartz. The fork-join queue and related systems with synchronization constraints: Stochastic ordering and computable bounds. *Advances in Applied Probability*, 21:629–660, 1989.

[7] 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. *Annals of Applied Probability*, 11:608–649, 2001.

[8] S. L. Bell and R. J. Williams. Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: Asymptotic optimality of a threshold policy. *Electronic Journal of Probability*, 10:1044–1115, 2005.

[9] D. Bertsimas, D. Gamarnik, and J. Tsitsiklis. Performance of multiclass Markovian queueing networks via piecewise linear Lyapunov functions. *Annals of Applied Probability*, 11(4):1384–1428, 2001.

[10] S. Bhadra, Y. Lu, and M. S. Squillante. Optimal capacity planning in stochastic loss networks with time-varying workloads. In *Proceedings of ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pages 227–238, New York, June 2007. ACM.

[11] T. Bonald. The Erlang model with non-Poisson call arrivals. In *Proceedings of Joint SIGMETRICS/Performance Conference on Measurement and Modeling of Computer Systems*, pages 276–286, New York, June 2006. ACM.

[12] O. J. Boxma, G. M. Koole, and Z. Liu. Queueing-theoretic solution methods for models of parallel and distributed systems. In O. J. Boxma and G. M. Koole, editors, *Performance Evaluation of Parallel and Distributed Systems*, pages 1–24. CWI Tract 105, Amsterdam, 1994.

[13] M. Bramson. Instability of FIFO queueing networks. *Annals of Applied Probability*, 4:414–431, 1994.

[14] H. Chen and D. D. Yao. Dynamic scheduling of a multiclass fluid network. *Operations Research*, 41(6):1104–1115, 1993.

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

[16] J. W. Cohen. *The Single Server Queue*. North Holland, First Edition, 1969. Second edition, 1982.

[17] J. W. Cohen. A two-queue, one-server model with priority for the longer queue. *Queueing Systems Theory and Applications*, 2(3):261–283, 1987.

[18] J. W. Cohen. Boundary value problems in queueing theory. *Queueing Systems Theory and Applications*, 3:97–128, 1988.

[19] J. W. Cohen and O. J. Boxma. *Boundary Value Problems in Queueing System Analysis*. North Holland, 1983.

[20] P. J. Courtois. *Decomposability*. Academic Press, 1977.

[21] P. J. Courtois and P. Semal. Error bounds for the analysis by decomposition of non-negative matrices. In *Proceedings of International Workshop on Applied Mathematics and Performance/Reliability Models of Computer/Communication Systems*, pages 253–268, 1983.

[22] J. G. Dai. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability*, 5:49–77, 1995.

[23] Y. Dallery and Y. Frein. On decomposition methods for tandem queueing networks with blocking. *Operations Research*, 41:386–399, 1993.

[24] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah. Throughput-delay trade-off in wireless networks. In *Proc. IEEE Infocom*, March 2004.

[25] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah. Optimal throughput-delay scaling in wireless networks - Part I: The fluid model. *IEEE Trans. Inform. Theory*, 52(6):2568–2592, 2006.

[26] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah. Optimal throughput-delay scaling in wireless networks - Part II: Constant-size packets. *IEEE Trans. Inform. Theory*, 52(11):5111–5116, 2006.

[27] G. Fayolle and R. Iasnogorodski. Two coupled processors: The reduction to a Reimann-Hilbert problem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47:325–351, 1979.

[28] G. Fayolle, R. Iasnogorodski, and V. Malyshev. *Random Walks in the Quarter-Plane: Algebraic Methods, Boundary Value Problems and Applications*. Springer-Verlag, 1999.

[29] G. Fayolle, V. A. Malyshev, and M. V. Menshikov. *Topics in the Constructive Theory of Countable Markov Chains*. Cambridge University Press, 1995.

[30] L. Flatto. The longer queue model. *Probability in the Engineering and Informational Sciences*, 3:537–559, 1989.

[31] L. Flatto and S. Hahn. Two parallel queues created by arrivals with two demands. *SIAM Journal on Applied Mathematics*, 44:1041–1053, 1984.

[32] L. K. Fleischer and J. Sethuraman. Efficient algorithms for separated continuous linear programs: The multicommodity flow problem with holding costs and extensions. *Mathematics of Operations Research*, 30(4):916–938, 2005.

[33] G. J. Foschini and J. Salz. A basic dynamic routing problem and diffusion. *IEEE Transactions on Communications*, COM-26(3):320–327, March 1978.

[34] D. Gamarnik and M. S. Squillante. Analysis of stochastic online bin packing processes. *Stochastic Models*, 21:401–425, 2005.

[35] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5:79–141, 2003.

[36] D. P. Gaver, Jr. A waiting line with interrupted service, including priorities. *Journal of the Royal Statistical Society, Series B*, 24:73–90, 1962.

[37] P. Gupta and P. R. Kumar. The capacity of wireless networks. *IEEE Trans. Inform. Theory*, 46(2):388–404, 2000.

[38] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29:567–588, 1981.

[39] M. Harchol-Balter, C. Li, T. Osogami, A. Scheller-Wolf, and M. S. Squillante. Cycle stealing under immediate dispatch task assignment. In *Proceedings of Annual ACM Symposium on Parallel Algorithms and Architectures*, pages 274–285, June 2003.

[40] M. Harchol-Balter, C. Li, T. Osogami, A. Scheller-Wolf, and M. S. Squillante. Task assignment with cycle stealing under central queue. In *Proceedings of International Conference on Distributed Computing Systems*, pages 628–637, May 2003.

[41] M. Harchol-Balter, T. Osogami, A. Scheller-Wolf, and A. Wierman. Multi-server queueing systems with multiple priority classes. *Queueing Systems Theory and Applications*, 51:331–360, 2005.

[42] J. M. Harrison. *Brownian Motion and Stochastic Flow Systems*. John Wiley and Sons, 1985.

[43] J. M. Harrison and M. J. Lopez. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems Theory and Applications*, 33:339–368, 1989.

[44] J. M. Harrison and A. Zeevi. Dynamic scheduling of a multiclass queue in the Halfin and Whitt heavy traffic regime. *Operations Research*, 52:243–257, 2004.

[45] N. K. Jaiswal. *Priority Queues*. Academic Press, 1968.

[46] P. Jelenković, P. Momčilović, and M. S. Squillante. Buffer scalability of wireless networks. In *Proc. IEEE Infocom*, April 2006.

[47] P. Jelenković, P. Momčilović, and M. S. Squillante. Scalability of wireless networks. *IEEE/ACM Trans. Networking*, 15(2), April 2007.

[48] A. Jovičić, P. Viswanath, and S. Kulkarni. Upper bounds to transport capacity of wireless networks. *IEEE Trans. Inform. Theory*, 50(11):2555–2565, 2004.

[49] K. Jung, Y. Lu, D. Shah, M. Sharma, and M. S. Squillante. Revisiting stochastic loss networks: Structures and algorithms. In *Proceedings of ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pages 407–418, New York, June 2008. ACM.

[50] A. Katok and B. Hasselblatt. *Introduction to the Modern Theory of Dynamical Systems*. Cambridge University Press, 1995.

[51] F. P. Kelly. Blocking probabilities in large circuit-switched networks. *Advances in Applied Probability*, 18(2):473–505, 1986.

[52] F. P. Kelly. Loss networks. *Annals of Applied Probability*, 1(3):319–378, 1991.

[53] P. R. Kumar. Re-entrant lines. *Queueing Systems Theory and Applications*, 13:87–110, 1993.

[54] P. R. Kumar and L.-L. Xie. A network information theory for wireless communications: Scaling laws and optimal operation. *IEEE Trans. Inform. Theory*, 50(5):748–767, 2004.

[55] H. J. Kushner and P. Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time*. Springer-Verlag, 1992.

[56] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM, Philadelphia, 1999.

[57] O. Leveque and E. Telatar. Information theoretic upper bounds on the capacity of ad hoc networks. *IEEE Trans. Inform. Theory*, 51(3):858–865, 2005.

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

[59] S. P. Meyn. Sequencing and routing in multiclass queueing networks. part I: Feedback regulation. *SIAM Journal of Control and Optimization*, 40:741–776, 2001.

[60] S. P. Meyn. Sequencing and routing in multiclass queueing networks. part II: Workload relaxations. *SIAM Journal of Control and Optimization*, 42:178–217, 2003.

[61] S. P. Meyn. Dynamic safety-stocks for asymptotic optimality in stochastic networks. *Queueing Systems Theory and Applications*, 50:255–297, 2005.

[62] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993. Available at probability.ca/MT.

[63] R. D. Nelson and T. K. Philips. An approximation for the mean response time for shortest queue routing with general interarrival and service times. *Performance Evaluation*, 17:123–139, 1993.

[64] R. D. Nelson and M. S. Squillante. Parallel-server stochastic systems with dynamic affinity scheduling and load balancing. Preprint, 2006.

[65] R. D. Nelson and A. N. Tantawi. Approximate analysis of fork/join synchronization in parallel queues. *IEEE Transactions on Computers*, 37(6):739–743, June 1988.

[66] M. F. Neuts. Moment formulas for the Markov renewal branching process. *Advances in Applied Probability*, 8:690–711, 1978.

[67] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. The Johns Hopkins University Press, 1981.

[68] M. F. Neuts. *Structured Stochastic Matrices of M/G/1 Type and Their Applications*. Marcel Dekker, 1989.

[69] T. Nowicki, M. S. Squillante, and C. W. Wu. Fundamentals of dynamic decentralized optimization in autonomic computing systems. In O. Babaoglu, M. Jelasity, A. Montresor, C. Fetzer, S. Leonardi, A. van Moorsel, and M. van Steen, editors, *Self-Star Properties in Complex Information Systems: Conceptual and Practical Foundations*, volume 3460 of *Lecture Notes in Computer Science*, pages 204–218. Springer-Verlag, 2005.

[70] T. Osogami. *Analysis of Multi-server Systems via Dimensionality Reduction of Markov Chains*. PhD thesis, Carnegie Mellon University, 2005.

[71] T. Osogami, M. Harchol-Balter, and A. Scheller-Wolf. Analysis of cycle stealing with switching times and thresholds. In *Proceedings of ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pages 184–195, New York, June 2003. ACM.

[72] T. Osogami, M. Harchol-Balter, and A. Scheller-Wolf. Analysis of cycle stealing with switching times and thresholds. *Performance Evaluation*, 61:347–369, 2005.

[73] T. Osogami, M. Harchol-Balter, A. Scheller-Wolf, and L. Zhang. Exploring threshold-based policies for load sharing. In *Proceedings of the 42nd Annual Allerton Conference on Communication, Control, and Computing*, 2004.

[74] V. G. Peris, M. S. Squillante, and V. K. Naik. Analysis of the impact of memory in distributed parallel processing systems. In *Proceedings of ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pages 5–18, New York, May 1994. ACM.

[75] L. Pontryagin, V. Boltyanskii, R. Gamkrelidze, and E. Mishchenko. *The Mathematical Theory of Optimal Processes*. Interscience Publishers, New York, 1962.

[76] D. Prener. Personal communication, 2007.

[77] M. I. Reiman and B. Simon. An interpolation approximation for queueing systems with Poisson input. *Operations Research*, 36(3):454–469, May-June 1988.

[78] A. N. Rybko and A. L. Stolyar. Ergodicity of stochastic processes describing the operations of open queueing networks. *Problems of Information Transmission*, 28:199–220, 1992.

[79] A. Seierstad and K. Sydsieter. Sufficient conditions in optimal control theory. *International Economic Review*, 18(2):367–391, June 1977.

[80] J. Sethuraman and M. S. Squillante. Optimal stochastic scheduling in multiclass parallel queues. In *Proceedings of ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pages 93–102, New York, June 1999. ACM.

[81] J. Sethuraman and M. S. Squillante. Analysis of parallel-server queues under spacesharing and timesharing disciplines. In G. Latouche and P. Taylor, editors, *Matrix-Analytic Methods: Theory and Applications*, pages 357–380. World Scientific, 2002.

[82] M. S. Squillante. *Issues in Shared-Memory Multiprocessor Scheduling: A Performance Analysis*. PhD thesis, Department of Computer Science, University of Washington, September 1990.

[83] M. S. Squillante. Stochastic analysis of resource allocation in parallel processing systems. In E. Gelenbe, editor, *Computer System Performance Modeling in Perspective: A Tribute to the Work of Prof. K.C. Sevcik*, pages 227–256. Imperial College Press, London, 2005.

[84] M. S. Squillante and K. P. Tsoukatos. Analysis of optimal scheduling in distributed parallel queueing systems. In *Proceedings of International Conference on Computer Communication*, August 1995.

[85] M. S. Squillante, F. Wang, and M. Papaefthymiou. Stochastic analysis of gang scheduling in parallel and distributed systems. *Performance Evaluation*, 27&28:273–296, October 1996.

[86] M. S. Squillante, Y. Zhang, A. Sivasubramaniam, and N. Gautam. Generalized parallel-server fork-join queues with dynamic task scheduling. *Annals of Operations Research*, 160:227–255, April 2008.

[87] M. S. Squillante, Y. Zhang, A. Sivasubramaniam, N. Gautam, H. Franke, and J. Moreira. Modeling and analysis of dynamic coscheduling in parallel and distributed environments. In *Proceedings of ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pages 43–54, New York, June 2002. ACM.

[88] G. W. Stewart. Computable error bounds for aggregated Markov chains. *Journal of the ACM*, 30:271–285, 1983.

[89] H. Takagi. *Queueing Analysis – A Foundation of Performance Evaluation*, volume 1: Vacation and Priority Systems, Part 1. North Holland, 1991.

[90] J. S. van Leeuwaarden, M. S. Squillante, and E. M. Winands. Quasi-birth-and-death processes, lattice path counting, and hypergeometric functions. *Journal of Applied Probability*, 46(2):507–520, June 2009.

[91] S. Varma and A. M. Makowski. Interpolation approximations for symmetric fork-join queues. *Performance Evaluation*, 20:245–265, 1994.

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

[93] W. Whitt. *Stochastic-Process Limits*. Springer-Verlag, New York, 2002.

[94] R. J. Williams. On dynamic scheduling of a parallel server system with complete resource pooling. *Fields Institute Communications*, 28:49–71, 2000.

[95] Y. Zheng and P. H. Zipkin. A queueing model to analyze the value of centralized inventory information. *Operations Research*, 38:296–307, 1990.

Mark S. Squillante
Mathematical Sciences Department
IBM Thomas J. Watson Research Center
P.O. Box 218, Yorktown Heights, NY 10598
USA
e-mail: mss@watson.ibm.com