Queueing-Theoretic Solution Methods for Models of Parallel and Distributed Systems

Onno Boxma[†], Ger Koole[†] & Zhen Liu[‡]

[†] *CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands*
[‡] *INRIA Sophia-Antipolis, B.P. 93, 06902 Sophia-Antipolis, France*

Abstract

This paper aims to give an overview of solution methods for the performance analysis of parallel and distributed systems. After a brief review of some important general solution methods, we discuss key models of parallel and distributed systems, and optimization issues, from the viewpoint of solution methodology.

AMS Subject Classification (1991): 60K20, 68M20
Keywords & Phrases: queueing, performance evaluation, parallel systems
Note: Supported by the European Grant BRA-QMIPS of CEC DG XIII, the second author is also supported by NFI.
This paper appeared also in *Performance Evaluation of Parallel and Distributed Systems — Solution Methods*, O.J. Boxma and G.M. Koole (eds.), CWI, Amsterdam, 1994 (CWI Tract 105 & 106).

1. Introduction

The purpose of this paper is to present a survey of queueing theoretic methods for the quantitative modeling and analysis of parallel and distributed systems. We discuss a number of queueing models that can be viewed as key models for the performance analysis and optimization of parallel and distributed systems. Most of these models are very simple, but display an essential feature of distributed processing. In their simplest form they allow an exact analysis. We explore the possibilities and limitations of existing solution methods for these key models, with the purpose of obtaining insight into the potential of these solution methods for more realistic complex quantitative models.

As far as references is concerned, we have restricted ourselves in the text mainly to key references that make a methodological contribution, and to surveys that give the reader further access to the literature; we apologize for any inadvertent omissions. The reader is referred to Gelenbe's book [65] for a general introduction to the area of multiprocessor performance modeling and analysis.

Stochastic Petri nets provide another formalism for modeling and performance analysis of discrete event systems. The reader is referred to the survey paper of Murata [127] for results of their qualitative analysis. The use of this tool for performance evaluation of parallel and distributed systems has recently become popular, as is illustrated in the special issue of *J. of Parallel and Distributed Computing* (Vol. 15, No. 3, July 1992). A survey on recent results of their quantitative analysis is found in the paper of Baccelli et al. [6].

Simulation is an important method for solving queueing models. In this paper, we will not discuss that approach. The interested reader is referred to the books of Mitrani [125], Sauer and MacNair [146] and Rubinstein [145].

The paper is organized as follows. Section 2 contains a global discussion of some solution methods that have been successful in the performance analysis of parallel and distributed systems. Six key models for this performance analysis are discussed in Section 3, with an emphasis on solution methodology. Section 4 is concerned with mathematical techniques for the optimal control of distributed systems. We distinguish between load balancing, routing, server allocation and scheduling.

2. Solution Methods

The publication of J.W. Cohen's 'The Single Server Queue' [37] in 1969 marked the end of an era in queueing theory, in which the emphasis in queueing research had been placed on the exact analysis of models with one server and/or one queue. Around 1970 successful applications of queueing theory to problems of computer performance began to appear. Rather simple queueing network models turned out to be able to yield quite accurate predictions of the behaviour of complex computer systems, thus stimulating queueing network research. Extensive queueing network results have been obtained in the seventies and eighties and have been made available for computer engineering purposes by the introduction of efficient numerical algorithms (see the surveys of Kleinrock [100] and Lavenberg [109]).

The performance analysis of parallel and distributed systems leads in a natural way to multidimensional queueing models. Generalization of the single-queue solution methods to those models is straightforward only in rare instances; and adaptation of the queueing network results to parallel and distributed systems is usually only possible by making gross simplifications. In this section we discuss a few solution methods that try to bridge this gap, and that have been successful in analyzing the performance of a number of (often admittedly simple) models of parallel and distributed systems: Product-form solutions, some methods from complex function theory, a number of analytic-algorithmic methods, heavy and light traffic approximations, the large deviation technique, and state recursions. The choice of these methods, above, e.g., aggregation and decomposition methods is undoubtedly influenced by the research interests of the authors. The field of performance analysis of parallel and distributed systems is still in its infancy, and it is yet far from clear which methods have the biggest potential for capturing the characteristic features of parallel systems.

2.1 Product-form solutions

An important contribution of queueing network theory is that, under certain assumptions, it allows one to obtain a simple exact solution for the joint queue length distribution in a separable form: the *product form* [96, 163, 158]. The reader is also referred to Liu & Nain [114] for recent extensions, efficient computational algorithms, and sensitivity analysis of product form queueing networks, and to Disney & König [50] for an extensive survey of queueing networks and their random processes.

Although the modeling of parallel and distributed systems only seldom leads to product forms, queueing networks do provide an important and widely used tool for modeling parallel

and distributed systems. We mention two interesting product-form applications. Heidelberger & Trivedi [83] present a class of parallel processing systems in which jobs subdivide in several asynchronous tasks; they approximate this non-product-form network iteratively by a sequence of product-form networks. An interesting new development in product-form theory has been stimulated by the performance analysis of resource request and allocation models with positive and negative signals; they were shown to give rise to product-form networks with positive and negative customers [66, 68]. The latter results are surveyed in [67]. See Harrison & Pitel [81] for some recent results and further references on *sojourn time* distributions in networks with positive and negative customers.

2.2 Methods from complex-function theory

The modeling of queueing systems with multiple queues and/or multiple servers frequently leads to multi-dimensional models with multiple unbounded components; in particular to the analysis of Markov processes whose state space is the N-dimensional set of lattice points with integer-valued non-negative coordinates. The functional equations arising in the analysis of such processes (obtained after taking transforms of, say, the joint queue length distribution) usually present formidable analytic difficulties.

For the two-dimensional case, however, techniques have been developed which often make it possible to reduce these functional equations to standard problems of the theory of boundary value equations (Wiener-Hopf, Dirichlet, Riemann, Riemann-Hilbert) and singular integral equations. Pioneering papers are those of Eisenberg [55] (transforming a two-queue polling problem into a Fredholm integral equation) and Fayolle & Iasnogorodski [57] (transforming a problem concerning two processors with coupled speeds into a Riemann-Hilbert boundary value problem). A systematic and detailed study of the 'boundary value method' is presented by Cohen & Boxma [43], with applications to various queueing problems: a two-queue polling problem, the shorter queue model, processors with coupled speeds, the M/G/2 queue. A concise exposition of the method, and several applications and references, are presented in [39]. A detailed investigation of random walks on the two-dimensional lattice in the first quadrant is continued by Cohen in [40]. This has led to a better understanding of, among others, the ergodicity conditions and the usefulness of the concept of (boundary) hitting points.

2.3 Analytic-algorithmic methods

The methods mentioned in the previous subsection are mainly applicable to some specific 'two-dimensional' models, and even then the performance measures are not always directly available. Therefore various analytic-algorithmic methods have been developed to solve multi-dimensional queueing systems. For multi-dimensional models with all but one component finite there are good analytic-algorithmic methods, like the well known *matrix geometric method* (Neuts [135]) and the related spectral method of Mitra & Mitrani [126]; see also the interesting and extensive methodological discussion by Gail et al. [64]. Below we discuss the power series algorithm and the compensation approach, two methods that are not yet so well known, that are mathematically interesting and that exploit the stochastic properties of queueing systems more than general methods based on, say, state space truncation and solving large systems of equations for Markov chains. It should be observed, though, that

recently much progress has been made in numerically handling large Markov chains. Several approaches are presented in the conference proceedings [150]; see also the survey of Grassmann [73].

The power series algorithm (introduced by Hooghiemstra et al. [85]) is a numerical procedure which can, formally, be applied to any Markov process (cf. [107]). It writes the stationary distribution of the process as a power series of some parameter, in queueing applications usually the load of the system. In a series of papers Blanc (with co-authors) has shown that the algorithm works well for many multi-dimensional queueing systems (see his survey [20]). However, the convergence properties and error estimates of the algorithm are still unknown, and therefore no guarantee can be given as to the convergence of the method for arbitrary models.

The compensation approach (developed in Adan's PhD thesis [1]) can be applied to two-dimensional homogeneous random walks in the first quadrant without transitions to the north, north-east and east. It writes the stationary distribution as a sum of product forms, which all satisfy the steady state equations on the interior, and where each additional term ensures, alternatively, that the steady state equations on the horizontal, respectively vertical, axis are satisfied. Generally the algorithm can be shown to converge exponentially fast. The algorithm is developed by Adan, Wessels & Zijm [2] for the shortest queue model. Other queueing models which are studied are a multiprogramming queue ([1], ch. 4), and the 2×2 clocked buffered switch of an interconnection network [26]. For the latter model, the method has been extended to a 3-dimensional case [159]. [26] indicates a link between the compensation approach and the boundary value method mentioned in subsection 2.2. Recent work of Cohen [41, 42] considerably adds to this insight. In [41] he studies the class of two-dimensional nearest neighbour random walks without transitions to the north, north-east and east, that is also considered in [1]. He shows that the bivariate generating function of the stationary distribution can be represented by a meromorphic function—an analytic function apart from a finite number of poles in every finite domain. The poles appear in fact as powers in the product forms in Adan's solution representation (cf. the correspondence between the representations $1/(1-az)$ and $\sum a^n z^n$). Cohen exposes the construction of this representation as a meromorphic function. He does this in much more detail for a special case of this class of random walks, the symmetrical shortest queue: he shows [42] how all poles, *and* all zeros, of the meromorphic generating function can be determined from the original functional equation. This leads to a simple expression for the main performance measures, which are easily calculated with any desired accuracy.

2.4 Heavy and light traffic approximations

Approximation techniques present an alternative approach to numerical methods for solving analytically intractable queueing systems. Heavy and light traffic approximations are among the most popular techniques of this kind.

By *heavy traffic* we mean that the system approaches saturation, so that the queues are nonempty most of the time. In this case, the queue lengths, when properly normalized, can be approximated by Brownian motions with drift, which leads to a diffusion approximation of the system. The reader is referred to the survey of Glynn [72]. For generalized Jackson queueing

networks, Harrison & Williams [78] prove the existence of stationary distributions of diffusions and their product form. They also show [79] that the notion of quasireversibility for queueing networks extends to the Brownian limit. Closed queueing networks are analyzed in Harrison, Williams & Chen [80]. Not all multiclass queueing networks can have such approximations (see e.g. [45]). Sufficient conditions for the existence and uniqueness of Brownian models are established in Reiman et al. [140] and Dai & Williams [46].

In *light traffic* approximations, a performance measure is considered as a function of the arrival rate. Derivatives of this function are computed at point zero. The light traffic approximations are developed in Burman & Smith [27, 28] for a single queue and in Reiman & Simon [139] for an open queueing network.

Approximations for moderate traffic can be obtained by interpolating heavy and light traffic approximations.

2.5 Large deviations

Although there is an extensive literature of large deviations on Markov processes (see Deuschel & Stroock [49], Dembo & Zeitouni [48]), it is only recently that these techniques have become important tools for solving queueing systems. Large deviations principle has been proved, under different statistical assumptions, for single queues, see e.g. Chang [31], Duffield & O'Connell [52], Liu et al. [116], and for queueing networks, see e.g. Tsoucas [157], Dupuis & Ellis [53].

A comprehensive treatment is provided in the forthcoming book [54]. Large deviations estimates have also been applied to rare event simulations, see Chang et al. [33].

2.6 State recursion

For queueing systems with synchronization constraints, as frequently occur in parallel systems, one can often write the dynamics in the form of a state recursion equation, generalizing Lindley's equation: Baccelli & Makowski [13], Baccelli & Liu [11]. Some of these state recursions lead to solvable integral equations as in the recent work of Jean-Marie [91]. Among the other techniques which were proposed based on state recursions, we would quote

- bounds: computable bounds on noncomputable stochastic models can often be obtained using stochastic ordering techniques (e.g. convex ordering, Schur convexity, association etc.). A good reference for this is the book by D. Stoyan [151]. A survey on the application of these techniques to synchronization problems can be found in [13].
- large deviation estimates as in Baccelli & Konstantopoulos [9]. This technique is based on the computation of the Cramer Legendre transform of the Perron Frobenius eigenvalue of the Laplace transform of the matrix that shows up in the state recursion.

3. Key Models

In this section we discuss some queueing models that can be viewed as key models for the performance analysis of parallel and distributed systems: fork-join, task graph, resequencing, shortest queue, polling and time warp models. The discussion is methodologically oriented, which has also guided our choice of references.

3.1 Fork-Join Model

The Fork-Join model is a simple queueing model of a parallel processing system. It consists of c parallel processors, each with a local queue. Each arriving job consists of c tasks, who each join the queue of a different processor (the fork primitive). A job is completed if all its tasks have completed service (the join primitive). Thus the model consists of c interrelated parallel queues (which are stochastically dependent due to the simultaneous arrivals).

For $c = 2$ and Poisson arrivals this model has been studied analytically. Flatto & Hahn [58] solve the model for inhomogeneous exponential servers, and obtain the limiting distribution as the number of tasks in one of the queues grows to infinity. This result is generalized by Wright [170], who also allows jobs consisting of a single task to join the system. Baccelli [5] solves the model for general, but exchangeable, service times using complex-function theory methods. De Klein [99] solves the model with general service times using the boundary value approach.

A completely different approach to obtain the asymptotic results is used by Shwartz & Weiss [148] (see also the afterword in [170]), who use large deviations and reversibility. Given there are n tasks in the second queue, they use reversibility to show that this queue built up with arrival and service rate reversed. It follows from the theory of large deviations that the moment at which the rates are reversed is almost deterministic, for n large. As the arrival rates of both queues are the same, conclusions can be drawn for the arrival rate in the first queue, and using transient results for the $M|M|1$ queue, the asymptotics are derived.

Because the analytic results only hold for $c = 2$ (and even then, performance measures are hard to obtain), attention has been paid to approximations and bounds. Both Nelson & Tantawi [130, 131] and Baccelli et al. [14] derive bounds for the system. In [131] bounds on the mean job response time are derived using inequalities on the maximum of associated random variables. In [14] the exponential conditions are dropped, and bounds on various performance measures are derived, again using associated random variables, but also stochastic orderings. It is interesting to note that both papers show that the response time grows logarithmically in the number of processors. Varma & Makowski [162] present an approximation for symmetric fork-join queues, interpolating between light-traffic and heavy-traffic results.

Kim & Agrawala [97] provide an algorithm to obtain the response times in the case of Erlang service time distributions.

Nelson et al. [133] compare the Fork-Join model with three other models with and without local queues and distributed processing. A central queue and distributed processing (which is equivalent to an $M|M|c$ queue with batch arrivals) performs best.

3.2 Task Graph Models

Directed acyclic graphs are frequently used to represent parallel programs, and are referred to as *task graphs*. In a monoprogramming system (i.e. a system where at most one parallel program runs at any time), the computation of program completion time can be performed by PERT techniques. We refer the reader to Baccelli et al. [8] for a survey on these techniques.

In order to model multiprogramming systems (i.e. systems where more than one parallel program can run simultaneously), queueing models with extended (synchronization) primi-

tives can be used.

The Fork-Join model can be considered as the ancestor of a line of models involving the execution of task graphs on parallel machines. Acyclic Fork-Join queueing models (corresponding to acyclic task graphs with precedence constraints) were introduced by Baccelli, Massey & Towsley in [15], where also a relation is indicated with the resequencing model that is considered below. The basic Fork-Join model is a special case of this acyclic case when all tasks have a single predecessor and a single successor.

A more general model with identical task graphs statically mapped on a set of processors was then introduced by Baccelli & Liu in [11]. The acyclic Fork-Join model mentioned above is a special case of this one when the number of processors is equal to the number of tasks in the graph. The model in [11] involves a non-trivial stability condition (which was recently understood in terms of so-called $(\max, +)$ Lyapunov exponents) and integral equations for the response times of tasks on processors that can be seen as the plain generalization of Lindley's integral equation. However, exact solutions of these equations are difficult except for very special cases.

Several variations on this basic model were proposed by the same authors in relation with distributed data base models ([10] and [113]). The main results for these models bear on (i) the shape of the stability region, (ii) the computation of the throughput either by stochastic ordering or using large deviation estimates as in [9], and (iii) bounds based on stochastic ordering. These models were also investigated using various asymptotic limits including the light traffic limits in the work of Varma [161] and diffusion limits as in the Stanford school around M. Harrison and in particular the work of Nguyen [136, 137].

3.3 Resequencing

The first stochastic resequencing models were infinite-server models, proposed in the context of reordering of packets in data communication networks. Kamoun, Kleinrock & Muntz [95] studied the exponential service time case using differential equations, and Baccelli, Gelenbe & Plateau [7] studied the non-exponential case using Wiener Hopf factorization. The model is basic in serialisation problems which arise quite naturally in various distributed algorithms. Models with no queueing effects were also considered by Harrus & Plateau [82] and by Varma [160], using analytical techniques. Recently Downey [51] made interesting new connections between the model considered in [7] and the cost of synchronization in parallel systems. The techniques used in these papers are mainly analytical ones based on complex variables, allowing to get simple series representations for the moments of response times etc.

In [166] Whitt presents a general exploration of overtaking phenomena in queueing networks; this includes an investigation of disordering in multiserver queues. He analyzes the number of jobs overtaken by an arbitrary job for $GI/M/s$ and $M/GI/s$ models with the First Come First Serve (FCFS) service policy. Iliadis & Lien [88] explicitly calculate the resequencing delay for two heterogeneous servers under two different threshold-type scheduling disciplines.

Other lines of thought consist in looking at

- more elaborate serialisation algorithms (e.g., timestamp ordering or two phase locking

[16], [13]). The analysis method is essentially that of state recursions.

- more structured disordering structures like interconnection networks as considered in the thesis of A. Jean-Marie [89] (see also [90]), where complex-analysis methods are used to compute the moments of the resequencing delays.

Resequencing is surveyed in [13].

3.4 The shortest queue and the smallest workload model

An example of a model with distributed processing is the *shortest queue model*. Here arriving customers join out of two (or more) queues the one with the least customers in it. Usually the arrival process is taken to be Poisson and the service times are taken exponential. The idea behind joining the shorter queue is that it *balances* the load in the system. For qualitative questions concerning this model, see section 4.2. Here we will deal with the quantitative aspects, i.e. with the *performance analysis* of the 2-queue shortest queue model.

Complex-variable methods have led to an exact analysis of the joint queue length process. Kingman [98] and Flatto & McKean [59] use a uniformization technique to determine the equilibrium distribution in the case of equal service rates. Fayolle and Iasnogorodski in their theses [56, 86] show that, even for asymmetric service rates, the problem can be reduced to a — generalized — Riemann-Hilbert boundary value problem; see also [43]. Knessl et al. [103] develop a scheme to obtain approximations for the joint queue length distribution, valid when one of the queue lengths is large. Foschini & Salz [60] employ a heavy traffic diffusion approximation. Interesting numerical approaches are proposed by Adan et al. [2] (the compensation approach), Blanc [19] (the power series algorithm, applicable to the case of more than two queues and general service times), Gertsbakh [71] (the matrix-geometric method) and Zhao & Grassmann [173] (who present an algorithm based on the results of [59]). Halfin [77] employs linear programming techniques to obtain bounds, and Nelson & Philips [134] present mean response time approximations for the case of K queues and general interarrival and service time distributions, assuming in their approximation method that the various queue lengths can differ by at most one.

For the 2-queue model where the customers are assigned to the queue with the smallest workload (and general service times) a performance analysis has been presented in [104]. Formal asymptotic approximations are constructed for the two-dimensional workload process, treating separately the asymptotic limits of heavy traffic, light traffic and large buffer contents. Cohen [36] presents an exact analysis of this M/G/2 queue, using a Wiener-Hopf decomposition. Cohen [38] also solves the 2-queue model with server priority for the longer queue (in a sense dual to the shortest queue model); here he uses a translation into a Riemann boundary value problem of a type that was not studied earlier in a queueing context.

3.5 Polling

The performance analysis of distributed systems often gives rise to single-server multi-queue *polling* models. The characteristic feature of polling models is that the server is moving between queues (which possibly requires switchover times), implying that the priority of the queues is dynamically (e.g., cyclically) changing. Some examples are token passing schemes in

local area networks with distributed channel access control, and resource arbitration and load sharing in multiprocessor computers. Many computer-communication examples of polling can be found in [74, 112, 153].

In a single-server cyclic polling model, the joint queue length process can — under some conditions on the service disciplines at the queues — be represented by a multi-type branching process with immigration [141]. The theory of such branching processes then immediately yields necessary and sufficient ergodicity conditions, and a complete solution for the joint queue length distribution. Unfortunately, the branching property does not hold for several important service disciplines, like those that put a limit on the number of services or the time of a server visit. In exceptional two-queue cases of the latter class, the joint queue length distribution can be determined by using the theory of Riemann-Hilbert boundary value problems [25, 43].

Proving ergodicity conditions for polling models generally is a challenging mathematical problem, for which recently considerable progress has been made; cf. the approach of [70] (based on stochastic dominance techniques and the well-known Loynes stability criteria for a queue in isolation), [4] (which uses Lyapunov functions for the verification of Foster's criterion), and [63] (based on a stochastic monotonicity property of the multidimensional queue length Markov chain at polling instants).

A quite generally valid result for (even non-cyclic) polling models is the pseudo-conservation law—an exact expression for a weighted sum of the mean queue lengths or mean waiting times [23]. The pseudo-conservation law has been extensively used to develop mean waiting time approximations.

Leung [110] has developed an interesting numerical procedure, based on the fast Fourier transform, that enables one in principle to determine polling performance measures with any required accuracy. The power series algorithm [18] is also applicable to a large class of polling models. An essential difficulty of these numerical techniques is their large computational complexity.

Takagi [152] gives an extensive bibliography of polling studies.

3.6 Time Warp

Simulations are usually well suited for parallel processing, especially if the physical model to be simulated consists of several components which can be simulated on different processors. Messages sent between the processors deal with the interaction between the components. A method to synchronize the components is the *Time Warp* protocol, as introduced by Jefferson [94]. Each processor continues the simulation, handling the already arrived messages. If a message arrives which should have been handled before, the processing *rolls back* to a point in time before the time associated with the message, and execution starts again. This mechanism can also be used for distributed systems other than simulation.

Besides local clocks for each component, there is a global clock, indicating a time before which no component has to be rolled back. The progression of the global time for specific models is the subject of several studies.

Kleinrock & Felderman [101] study a discrete-time model with two processors. The local

times of the processors increase with geometric jumps and sojourns. After each jump a message for the other processor is generated with a fixed probability. If that processor's local time is ahead of the time of the message, then it rolls back to that time. A related Markov chain is studied and the speed-up, relative to a single processor, is calculated. The results of [101] are a superset of those of Lavenberg et al. [108]. Mitra & Mitrani [124] analyze a model related to that of [101] in which the jump sizes can be arbitrary; their approach is based on a Wiener-Hopf factorization.

In Akyildiz et al. [3] a model with c processors and a limited shared memory capacity is analyzed using a simple Markov process, which approximates the used memory space. The results are compared with experimental data.

4. Optimization

In this section, we discuss optimization issues of parallel and distributed systems which can be tackled using queueing network formalisms. We shall first provide a general discussion about load balancing problems. Then we discuss in more detail the routing problem which is a special case of the load balancing problem, followed by a discussion on a dual problem, the problem of server allocation. In the last subsection, we will consider scheduling problems.

4.1 Load balancing

An operational aspect of distributed systems is the availability of a protocol which optimally balances the workload over the servers: a *load balancing protocol*. These protocols can roughly be divided into routing models, where at their moment of arrival in the system jobs are (irrevocably) routed to one of the servers, and server allocation models, where the servers determine from which input sources they draw their jobs.

Another important element of a load balancing protocol is the information it requires to operate. This information can range from total knowledge about the system at any point in time, to only information about some basic characteristics, like arrival rate and service times. In general, the term *dynamic* is used for policies which operate under time dependent information, whereas protocols operating under time independent characteristics of the system are called *static*. Below we give overviews of routing, server allocation and stochastic scheduling models, again with an emphasis on methodology. See Gelenbe & Pekergin [69] for an interesting general discussion on load balancing in parallel and distributed systems, that also touches upon the trade-off between static and dynamic load balancing; see Wang & Morris [164] for a taxonomy of the current load balancing protocols, discriminating between routing (called *source initiative*) and server allocation (*server initiative*) models. They provide numerical comparisons, based on analysis and simulation, of various allocation protocols, both static and dynamic.

4.2 Routing

The routing or customer allocation problem, as a special case of the load balancing problem, consists in assigning arriving customers to one of several parallel queues (which are usually assumed to have a single server). Thus, no jockeying amongst the queues is allowed. We will consider both static and dynamic routing problems.

For all sorts of information structure both the symmetric (i.e., the service times are equally distributed for each queue) and the asymmetric case are studied. For the symmetric models it is often possible to find the optimal policy, mostly using coupling, dynamic programming or stochastic orderings. Consequently, these are transient results, which often hold for a large class of cost functions and general arrivals.

Asymmetric models on the other hand rarely have a simple optimal policy; it usually depends on the arrival process, the service times, etc. The analysis is therefore often numerical in nature, Poisson arrivals are assumed and only long-run results are obtained.

Two static allocation policies have been proposed: probabilistic allocation (assign arriving jobs to a queue according to a fixed probability), and pattern allocation (route arriving jobs to a queue according to a routing table).

When the servers are identical the symmetric (or equal probability) routing policy is optimal among the probabilistic policies for the minimization of response times and resequencing delay. This can be shown using stochastic orderings and coupling (Chang et al. [32], Gün & Jean-Marie [92]).

For the general non-symmetrical problem, Buzen & Chen [30] present an algorithm for determining the probabilistic allocation which minimizes the mean sojourn time of a job.

In most of the numerical studies, queueing theory is used to determine an expression for the performance measure that is to be minimized. The separability of that expression in terms relating to only one particular queue, and the convexity of each term, lead to a tractable non-linear optimization problem of the class of resource allocation problems that is extensively discussed in the book of Ibaraki & Katoh [87].

Ross & Yao [144] study a probabilistic allocation problem with additional dedicated arrival streams and local priority scheduling. Proving convexity in their case is an interesting problem, that is solved using matroid theory. Bonomi & Kumar [22] also discuss probabilistic allocation with additional dedicated arrival streams. They consider the situation where not all system parameters are known, or where some of the parameters may change from time to time. They propose several *adaptive* load balancing algorithms, using stochastic approximation and stochastic control methods.

Pattern allocation leads to a more regular arrival process than probabilistic allocation, and hence better performance can be expected. However, constructing the optimal pattern is generally an unsolved problem. Various studies have been carried out for characterizing the optimal routing policies, see for example [17, 143] and the references therein.

When the servers are identical, Walrand [163] uses coupling arguments to show that assigning the jobs cyclicly to the queues (the round robin policy) is optimal for exponential service times. Recently, using a coupling technique and majorization theory, Liu & Towsley [119] generalized the optimality of the round robin policy to the case of identical IFR (Increasing Failure Rate) servers.

Under the assumption of general service time distributions, the round robin policy yields smaller (in the sense of increasing convex ordering) stationary and transient job waiting times than the symmetric probabilistic routing policy (Stoyan [151], Jean-Marie & Liu [93]).

For the case of non-identical servers Hajek [76] proves how the pattern allocation to a single queue should be, given that a fixed fraction of the arrivals should be sent to that queue, to minimize the average number of customers in that queue. His proof involves showing *multimodularity* (a generalization of convexity to multiple dimensions) of certain functions. Ramakrishnan [138] proposes a useful approximation procedure for non-identical exponential servers; see [44] for the case of general servers. Again separability of the objective function and convexity of each term are exploited.

A well studied dynamic model is that with exponential servers and decisions based on the numbers of jobs in the queues. For the symmetric model the optimality of shortest queue routing was first proved by Winston [167]. This result has been generalized in different directions by various authors. The techniques used are dynamic programming and coupling. A recent paper, showing the optimality for Schur convex cost functions, ILR (i.e., increasing in likelihood ratio) service time distributions and including finite buffers, is [156]. For asymmetric models the optimal policy does not have a nice structure. Several authors have tried to obtain good policies (e.g., Shenker & Weinrib [147]). Using dynamic programming, Hajek [75] showed for the model with two queues (and some additional features) that there is a non-decreasing switching curve. Xu & Chen [171] considered the limiting behavior of this curve for discounted costs, and showed that it converges to a constant, for unequal holding cost rates.

A model with a different information structure is the one where decisions are based on the *workload* in the queues. Routing to the queue with the smallest workload minimizes both the total workload (in fact, each weak Schur convex function of the workload vector is minimized) and the job response times. Note that the smallest workload policy is equivalent to FCFS. Some references are Wolff [168, 169], Foss [61, 62] and Daley [47]. It is interesting to note that basically all results are established using the same coupling argument. A generalization to network models, and an extensive list of references, can be found in [105].

An overview of routing policies and their performances is given by Boel & Van Schuppen [21]. They consider the problem from a control point of view, and discuss the question what amount of information is required at the routing points to achieve good system performance. Their paper concentrates on analytically and numerically tractable models.

4.3 Server allocation

As a dual problem to routing problems (which can be seen as job allocation problems), the problem of server allocation has also received much interest in the literature. However, there are few results on static server allocation problems.

We will restrict ourselves to single server models, one reason being that the results on multiple server models are less relevant to the present survey, the other being that policies, which are optimal for single server models, often perform very well if applied to multiple server models (e.g., Weiss [165]). As a general reference for multiple server models, we refer to Righter [142].

In the simplest model jobs arrive in several queues (single queue models are discussed in Subsection 4.3 on scheduling), each requiring an exponentially distributed amount of processing, depending on the queue. The objective is to minimize the weighted holding costs. Using

a simple interchange argument it can be shown that the single server should process the jobs (preemptively) in decreasing order of product of processing rate and holding cost rate [29]. This policy is known as the μc or $c\mu$ rule. Such a type of policy is called a *list policy*, i.e., a policy which has associated a list of the queues, and which processes the job whose queue is highest in the list.

Generalizations are possible in several directions. Assume that the service times are general, and that jobs, after completing service, can re-enter in a, possibly different, queue. We restrict ourselves to Poisson arrivals and non-preemptive policies. The problem of finding the server allocation policy that minimizes the weighted number of jobs in the system is known as *Klimov's problem*. It is shown in [102] that a list policy is optimal. A good reference for Klimov's and related problems is chapter 9 in [163].

If we assume that there are switching times between serving different queues, we arrive at polling models (cf. Subsection 2.5). Polling optimization issues have only recently been tackled. Some static and dynamic server routing optimization problems are reviewed in [24] and [172], respectively. Symmetric optimization models are discussed in [115], and an asymmetric optimization model is studied in [106].

Contrary to most studies discussed so far, the optimal policy in the model of Menich & Serfozo [123] is not a list policy. They augment a symmetric routing model with a movable server, and show that it should be assigned to the longest queue. In the case of finite buffers, the duality between various job allocation and server allocation problems where queue lengths are available to the controller has been established in Sparaggis et al. [149].

4.4 Scheduling

In most routing or server allocation models jobs are processed in FCFS order. Here this and other service policies are discussed. By scheduling we mean policies that determine the order according to which servers serve jobs waiting in the queue.

Consider a single $G/GI/s$ queue. Then the FCFS policy minimizes the stationary waiting times in the sense of the increasing convex ordering, in the case that the service time distribution is of IFR type (Hirayama & Kijima [84], Chang & Yao [35]).

Now assume that every job has a due date. Several papers have studied the effect that different scheduling policies have on the job lateness (defined as the amount of time the completion time of a job exceeds the due date of that job). The optimality of stochastic versions of SDD (the policy which processes jobs with the shortest due date first) has been established in Liu & Towsley [118]. Also for the $G/M/s$ queue when jobs have hard deadlines (meaning that a job leaves the system either when it finishes service or when its due date occurs) SDD is optimal (Towsley & Panwar [155]).

For queueing networks, where each queue has its own servers, the optimality of SDD was first shown in Towsley & Baccelli [154] for queues in tandem in the sense of convex ordering. More general results were established by Liu & Towsley [120] for in-forest networks consisting of multi-server queues. In [120], extremal properties of FCFS, LCFS (Last Come First Serve), stochastic SDD and LDD (standing for longest due date) policies were proved for the minimization (or maximization) of job response times, lateness and end-to-end delays.

The scheduling problem in more realistic parallel processing models has been addressed in a recent paper by Baccelli, Liu & Towsley [12]. They provide extremal properties of various scheduling strategies for multiprogrammed multitasked multiprocessor systems where task executions are constrained by precedence relations.

In most of the above mentioned studies, the techniques used are stochastic comparison and sample path analysis. Quite strong stochastic qualitative properties are established when these techniques are applicable. A general and unified theoretical formalism of these techniques has been proposed in a recent paper of Liu et al. [117].

In the following studies, the authors use queueing analysis to compare average performance measures of different scheduling policies.

Nelson et al. [34, 121, 122] considered the problem of allocating parallel tasks to processors so as to minimize job (consisting of parallel tasks) response times. They showed that allocation of tasks to different processors is not always a good strategy.

Performance analysis of scheduling policies in multiprogrammed multiprocessor systems with parallel tasks can be found in Nelson et al. [133], [128] for FCFS policies, in Towsley et al. [129] for processor sharing, and in Nelson & Towsley [132] for priority policies. Leutenegger & Vernon [111] provide a comparison of performances of various scheduling policies.

References

[1] I.J.B.F. Adan. *A Compensation Approach for Queueing Problems*. PhD thesis, Eindhoven University of Technology, 1991.

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

[3] I.F. Akyildiz, L. Chen, S.R. Das, R.M. Fujimoto, and R.F. Serfozo. Performance analysis of "time warp" with limited memory. *Performance Evaluation Review*, 20:213–224, 1992.

[4] E. Altman, P. Konstantopoulos, and Z. Liu. Stability, monotonicity and invariant quantities in general polling systems. *Queueing Systems*, 11:35–57, 1992.

[5] F. Baccelli. Two parallel queues created by arrivals with two demands: The $M/G/2$ symmetrical case. Technical report 426, INRIA-Rocquencourt, 1985.

[6] F. Baccelli, G. Balbo, R.J. Boucherie, J. Campos, and G. Chiola. Annotated bibliography on stochastic Petri nets. In O.J. Boxma and G.M. Koole, editors, *Performance Evaluation of Parallel and Distributed Systems — Solution Methods*. CWI, Amsterdam, 1994. CWI Tract 105 & 106.

[7] F. Baccelli, E. Gelenbe, and B. Plateau. An end-to-end approach to the resequencing problem. *Journal of the ACM*, 31:474–485, 1984.

[8] F. Baccelli, A. Jean-Marie, and Z. Liu. A survey on solution methods for task graph models. In N. Götz, U. Herzog, and M. Rettelbach, editors, *Second QMIPS Workshop*, volume 26(14) of *Arbeitsberichte der IMMD*, Erlangen, 1993.

[9] F. Baccelli and T. Konstantopoulos. Estimates of cycle times in stochastic Petri nets. In I. Karatzas and D. Ocone, editors, *Applied Stochastic Analysis*, pages 1–20. Springer Verlag, 1992. Lecture Notes in Control and Information Sciences 177.

[10] F. Baccelli and Z. Liu. On the stability condition of a precedence-based queueing discipline. *Journal of Applied Probability*, 21:883–898, 1989.

[11] F. Baccelli and Z. Liu. On the execution of parallel programs on multiprocessor systems—a queueing theory approach. *Journal of the ACM*, 37:373–414, 1990.

[12] F. Baccelli, Z. Liu, and D. Towsley. Extremal scheduling of parallel processing with and without real-time constraints. *Journal of the ACM*, 40:1209–1237, 1993.

[13] F. Baccelli and A.M. Makowski. Synchronization in queueing systems. In H. Takagi, editor, *Stochastic Analysis of Computer and Communication Systems*, pages 57–129. North-Holland, Amsterdam, 1990.

[14] 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.

[15] F. Baccelli, W.A. Massey, and D. Towsley. Acyclic fork-join queuing networks. *Journal of the ACM*, 36:615–642, 1989.

[16] F. Baccelli and Ph. Robert. Analysis of update response times in a distributed database maintained by the conservative time stamps ordering algorithm. In *Performance '83*, pages 415–436, 1983.

[17] C.E. Bell and S. Stidham, Jr. Individual versus social optimization in the allocation of customers to alternative servers. *Management Science*, 29:831–839, 1983.

[18] J.P.C. Blanc. A numerical approach to cyclic-service queueing models. *Queueing Systems*, 6:173–188, 1990.

[19] J.P.C. Blanc. The power-series algorithm applied to the shortest-queue model. *Operations Research*, 40:157–167, 1992.

[20] J.P.C. Blanc. Performance analysis and optimization with the power-series algorithm. In L. Donatiello and R.D. Nelson, editors, *Performance Evaluation of Computer and Communication Systems*, pages 53–80. North-Holland, Amsterdam, 1993.

[21] R.K. Boel and J.H. van Schuppen. Distributed routing for load balancing. *Proceedings of the IEEE*, 77:210–221, 1989.

[22] F. Bonomi and S. Kumar. Adaptive optimal load balancing in a nonhomogeneous multiserver system with a central job scheduler. *IEEE Transactions on Computers*, 39:1232–1250, 1990.

[23] O.J. Boxma. Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems*, 5:185–214, 1989.

[24] O.J. Boxma. Analysis and optimization of polling systems. In J.W. Cohen and C.D. Pack, editors, *Queueing, Performance and Control in ATM*, pages 173–183. North-Holland, Amsterdam, 1991.

[25] O.J. Boxma and W.P. Groenendijk. Two queues with alternating service and switching times. In O.J. Boxma and R. Syski, editors, *Queueing Theory and its Applications*, pages 261–282. North-Holland, Amsterdam, 1988.

[26] O.J. Boxma and G.J. van Houtum. The compensation approach applied to a 2×2 switch. *Probability in the Engineering and Informational Sciences*, 7:171–193, 1993. Reprinted in these proceedings.

[27] D. Y. Burman and D. R. Smith. A light traffic theorem for multi-server queues. *Mathematics of Operations Research*, 8:15–25, 1983.

[28] D. Y. Burman and D. R. Smith. An asymptotic analysis of a queueing system with Markov-modulated arrivals. *Operations Research*, 34:105–119, 1986.

[29] C. Buyukkoc, P. Varaiya, and J. Walrand. The $c\mu$ rule revisited. *Advances in Applied Probability*, 17:237–238, 1985.

[30] J.P. Buzen and P.P.-S. Chen. Optimal load balancing in memory hierarchies. In J.L. Rosenfeld, editor, *Proceedings of IFIP*, pages 271–275. North-Holland, Amsterdam, 1974.

[31] C.-S. Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, to appear 1994.

[32] C.-S. Chang, X.L. Chao, and M. Pinedo. A note on queues with Bernoulli routing. In *Proc. 29th Conf. on Decision and Control*, December 1990.

[33] C.-S. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin. The application of effective bandwidth to fast simulation of communication networks. *ACM/IEEE Trans. on Networking*, to appear 1994.

[34] C.-S. Chang, R. Nelson, and D.D. Yao. Optimal task scheduling on distributed parallel processors. Technical report, IBM Research Division, Yorktown Heights, 1992.

[35] C.-S. Chang and D.D. Yao. Rearrangement, majorization and stochastic scheduling. Technical report RC 16250, IBM Research Division, Yorktown Heights, 1990.

[36] J.W. Cohen. On the M/G/2 queueing model. *Stochastic Processes and their Applications*, 12:231–248, 1982.

[37] J.W. Cohen. *The Single Server Queue*. North-Holland, Amsterdam, 1982.

[38] J.W. Cohen. A two-queue, one-server model with priority for the longer queue. *Queueing Systems*, 2:261–283, 1987.

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

[40] J.W. Cohen. *Analysis of Random Walks*. IOS Press, Amsterdam, 1992.

[41] J.W. Cohen. On a class of two-dimensional nearest neighbouring random walks. *Journal of Applied Probability*, 1994. To appear.

[42] J.W. Cohen. On the analysis of the symmetrical shortest queue. Technical report BS-R9420, CWI, Amsterdam, 1994.

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

[44] M.B. Combé and O.J. Boxma. Optimization of static traffic allocation policies. *Theoretical Computer Science*, 125:17–43, 1994.

[45] J. Dai and Y. Wang. Nonexistence of Brownian models of certain multiclass queueing networks. *Queueing Systems*, 13:41–46, 1993.

[46] J. Dai and R. J. Williams. Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons. Working paper, 1994.

[47] D.J. Daley. Certain optimality properties of the first-come first-served discipline for $G|G|s$ queues. *Stochastic Processes and their Applications*, 25:301–308, 1987.

[48] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Jones and Bartlett Publishers, Boston, 1993.

[49] J.-D. Deuschel and D. W. Stroock. *Large Deviations*. Academic Press, San Diego, 1989.

[50] R. L. Disney and D. König. Queueing networks: A survey of their random processes. *SIAM Review*, 27:335–403, 1985.

[51] P. Downey. Bounds and approximations for overheads in the time to join parallel forks. Working paper, 1994.

[52] N. G. Duffield and N. O'Connell. Large deviations and overflow probabilities for the general single-server queue, with applications. Technical report DIAS-STP-93-30, Dublin City University, 1993.

[53] P. Dupuis and R. S. Ellis. The large deviation principle for a general class of queueing systems, I. Technical report LCDS 93-10, Brown University, 1993.

[54] P. Dupuis and R. S. Ellis. A weak convergence approach to the theory of large deviations. Technical report LCDS 93-6, Brown University, 1993. This manuscript will be published as a book by J. Wiley & Sons.

[55] M. Eisenberg. Two queues with alternating service. *SIAM Journal on Applied Mathematics*, 36:287–303, 1979.

[56] G. Fayolle. *Methodes Analytiques pour les Files d'Attente Couplees*. PhD thesis, Université de Paris VI, Paris, 1979.

[57] G. Fayolle and R. Iasnogorodski. Two coupled processors: The reduction to a Riemann-Hilbert problem. *Z. Wahrsch. Verw. Gebiete*, 47:325–351, 1979.

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

[59] L. Flatto and H.P. McKean. Two queues in parallel. *Comm. Pure Appl. Math.*, 30:255–263, 1977.

[60] G.J. Foschini and J. Salz. A basic dynamic routing problem and diffusion. *IEEE Transactions on Communications*, 26:320–327, 1978.

[61] S.G. Foss. Approximation of multichannel queueing systems. *Siberian Mathematical Journal*, 21:851–857, 1981.

[62] S.G. Foss. Comparison of servicing strategies in multichannel queueing systems. *Siberian Mathematical Journal*, 22:142–147, 1981.

[63] Ch. Fricker and R. Jaibi. Monotonicity and stability of periodic polling models. *Queueing Systems*, 15:211–238, 1994.

[64] H.R. Gail, S.L. Hantler, and B.A. Taylor. Spectral analysis of $M/G/1$ type Markov chains. Technical report, IBM Research Division, Yorktown Heights, 1992.

[65] E. Gelenbe. *Multiprocessor Performance*. Wiley, New York, 1989.

[66] E. Gelenbe. Product-form queueing networks with negative and positive customers. *Journal of Applied Probability*, 28:656–663, 1991.

[67] E. Gelenbe. G-networks: A unifying model for neural and queueing networks. *Annals of Operations Research*, 48:433–461, 1994. (Special issue on queueing networks, ed. N.M. van Dijk).

[68] E. Gelenbe, P. Glynn, and K. Sigman. Queues with negative arrivals. *Journal of Applied Probability*, 28:245–250, 1991.

[69] E. Gelenbe and F. Pekergin. Load balancing pragmatics. Technical report, EHEI Université René Descartes, February 1993.

[70] L. Georgiadis and W. Szpankowski. Stability of token passing rings. *Queueing Systems*, 11:7–33, 1992.

[71] I. Gertsbakh. The shorter queue problem: A numerical study using the matrix geometric solution. *European Journal on Operations Research*, 15:374–381, 1984.

[72] P. W. Glynn. Diffusion approximations. In D.P. Heyman and M.J. Sobel, editors, *Handbook on Operations Research and Management Science, Vol. 2*, pages 145–198. Elsevier Science Publishers B.V., Amsterdam, 1990.

[73] W.K. Grassmann. Computational methods in probability theory. In D.P. Heyman and M.J. Sobel, editors, *Handbook on Operations Research and Management Science, Vol. 2*, pages 199–254. Elsevier Science Publishers B.V., Amsterdam, 1990.

[74] D. Grillo. Polling mechanism models in communication systems – some application examples. In H. Takagi, editor, *Stochastic Analysis of Computer and Communication Systems*, pages 659–698. North-Holland, Amsterdam, 1990.

[75] B. Hajek. Optimal control of two interacting service stations. *IEEE Transactions on Automatic Control*, 29:491–499, 1984.

[76] B. Hajek. Extremal splittings of point processes. *Mathematics of Operations Research*, 10:543–556, 1985.

[77] S. Halfin. The shortest queue problem. *Journal of Applied Probability*, 22:865–878, 1985.

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

[79] J. M. Harrison and R. J. Williams. On the quasireversibility of a multiclass Brownian service station. Working paper, 1989.

[80] J. M. Harrison, R. J. Williams, and H. Chen. Brownian models of closed queueing networks with homogeneous customer populations. *Stochastics*, 29:37–74, 1990.

[81] P.G. Harrison and E. Pitel. Response time distributions in tandem G-networks. In O.J. Boxma and G.M. Koole, editors, *Performance Evaluation of Parallel and Distributed Systems — Solution Methods*. CWI, Amsterdam, 1994. CWI Tract 105 & 106.

[82] G. Harrus and B. Plateau. Queueing analysis of a reordering issue. *IEEE Transactions on Software Engineering*, 8:113–122, 1982.

[83] P. Heidelberger and K.S. Trivedi. Queueing network models for parallel processing with asynchronous tasks. *IEEE Transactions on Computers*, 31:1099–1109, 1982.

[84] T. Hirayama and M. Kijima. An extremal property of FIFO discipline in $G/IFR/1$ queues. *Advances in Applied Probability*, 21:481–484, 1989.

[85] G. Hooghiemstra, M. Keane, and S. van de Ree. Power series for stationary distributions of coupled processor models. *SIAM Journal on Applied Mathematics*, 48:1159–1166, 1988.

[86] R. Iasnogorodski. *Problèmes-Frontières dans les Files d'Attente*. PhD thesis, Université de Paris VI, Paris, 1979.

[87] T.I. Ibaraki and N. Katoh. *Resource Allocation Problems*. MIT Press, Cambridge, 1988.

[88] I. Iliadis and L.Y.-C. Lien. Resequencing delay for a queueing system with two heterogeneous servers under a threshold-type scheduling. *IEEE Transactions on Communications*, 36:692–702, 1988.

[89] A. Jean-Marie. *Modélisation de réseaux d'interconnexion multi etages*. PhD thesis, Université Paris Sud, 1987.

[90] A. Jean-Marie. Load balancing in a system of two queues with resequencing. In P.J. Courtois and G. Latouche, editors, *Performance '87*, pages 75–88. North-Holland, Amsterdam, 1988.

[91] A. Jean-Marie. Analytic computation of Lyapunov exponents in stochastic event graphs. In O.J. Boxma and G.M. Koole, editors, *Performance Evaluation of Parallel and Distributed Systems — Solution Methods*. CWI, Amsterdam, 1994. CWI Tract 105 & 106.

[92] A. Jean-Marie and L. Gün. Parallel queues with resequencing. *Journal of the ACM*, 40:1188–1208, 1993.

[93] A. Jean-Marie and Z. Liu. Stochastic comparisons for queueing models via random sums and intervals. *Advances in Applied Probability*, 24:960–985, 1992.

[94] D.R. Jefferson. Virtual time. *ACM Transactions on Programming Languages and Systems*, 7:404–425, 1985.

[95] F. Kamoun, L. Kleinrock, and R. Muntz. Queueing analysis of a reordering issue in the distributed database concurrency control mechanism. In *Proceedings of the 2nd International Conference on Distributed Computing Systems*, pages 13–23, 1982.

[96] F.P. Kelly. *Reversibility and Stochastic Networks*. Wiley, New York, 1979.

[97] C. Kim and A.K. Agrawala. Analysis of the fork-join queue. *IEEE Transactions on Computers*, 38:250–255, 1989.

[98] J.F.C. Kingman. Two similar queues in parallel. *Annals of Mathematical Statistics*, 32:1314–1323, 1961.

[99] S.J. de Klein. *Fredholm Integral Equations in Queueing Analysis*. PhD thesis, University of Utrecht, 1988.

[100] L. Kleinrock. Performance evaluation of distributed computer-communication systems. In O.J. Boxma and R. Syski, editors, *Queueing Theory and its Applications*, pages 1–57. North-Holland, Amsterdam, 1988.

[101] L. Kleinrock and R.E. Felderman. Two processor time warp analysis: A unifying approach. *International Journal in Computer Simulation*, 2:345–371, 1992.

[102] G.P. Klimov. Time-sharing service systems I. *Theory of Probability and its Applications*, 19:532–551, 1974.

[103] C. Knessl, B.J. Matkowsky, Z. Schuss, and C. Tier. Two parallel queues with dynamic routing. *IEEE Transactions on Communications*, 34:1170–1176, 1986.

[104] C. Knessl, B.J. Matkowsky, Z. Schuss, and C. Tier. Two parallel $M/G/1$ queues where arrivals join the system with the smaller buffer content. *IEEE Transactions on Communications*, 35:1153–1158, 1987.

[105] G.M. Koole. On the optimality of FCFS for networks of multi-server queues. Technical report BS-R9235, CWI, Amsterdam, 1992.

[106] G.M. Koole. Assigning a single server to inhomogeneous queues with switching costs. Technical report BS-R9405, CWI, Amsterdam, 1994.

[107] G.M. Koole. On the power series algorithm. In O.J. Boxma and G.M. Koole, editors, *Performance Evaluation of Parallel and Distributed Systems — Solution Methods*. CWI, Amsterdam, 1994. CWI Tract 105 & 106.

[108] S. Lavenberg, R. Muntz, and B. Samadi. Performance analysis of a rollback method for distributed simulation. In *Performance '83*, pages 117–132, 1983.

[109] S.S. Lavenberg. A perspective on queueing models of computer performance. In O.J. Boxma and R. Syski, editors, *Queueing Theory and its Applications*, pages 59–94. North-Holland, Amsterdam, 1988.

[110] K.K. Leung. Waiting time distributions for token-passing systems with limited-k service via discrete Fourier transforms. In P.J.B. King, I. Mitrani, and R.J. Pooley, editors, *Performance '90*, pages 333–347. North-Holland, Amsterdam, 1990.

[111] S.T. Leutenegger and M.K. Vernon. The performance of multiprogrammed multiprocessor scheduling policies. Technical Report TR 913, University of Wisconsin-Madison, 1990.

[112] H. Levy and M. Sidi. Polling systems: Applications, modeling and optimization. *IEEE Transactions on Communications*, 38:1750–1760, 1990.

[113] Z. Liu and F. Baccelli. Generalized precedence-based queueing systems. *Mathematics of Operations Research*, 17:615–639, 1992.

[114] Z. Liu and P. Nain. Sensitivity results in open, closed, and mixed product-form queueing networks. *Performance Evaluation*, 13:237–251, 1991.

[115] Z. Liu, P. Nain, and D. Towsley. On optimal polling policies. *Queueing Systems*, 11:59–83, 1992.

[116] Z. Liu, P. Nain, and D. Towsley. Exponential bounds with an application to call admission. Technical report, University of Massachusetts, 1994.

[117] Z. Liu, P. Nain, and D. Towsley. Sample path methods in the control of queues. *Queueing Systems*, 1994. To appear.

[118] Z. Liu and D. Towsley. Effects of service disciplines in $G/G/s$ queueing systems. *Annals of Operations Research*, 48:401–429, 1994. (Special issue on queueing networks, ed. N.M. van Dijk).

[119] Z. Liu and D. Towsley. Optimality of the round robin routing policy. *Journal of Applied Probability*, 1994. To appear.

[120] Z. Liu and D. Towsley. Stochastic scheduling in in-forest networks. *Advances in Applied Probability*, 26:222–241, 1994.

[121] A.M. Makowski and R. Nelson. Distributed parallelism considered harmful. Technical report, IBM Research Division, Yorktown Heights, 1991.

[122] A.M. Makowski and R. Nelson. Optimal scheduling for a distributed parallel processing model. Technical report, IBM Research Division, Yorktown Heights, 1992.

[123] R. Menich and R.F. Serfozo. Optimality of routing and servicing in dependent parallel processing systems. *Queueing Systems*, 9:403–418, 1991.

[124] D. Mitra and I. Mitrani. Analysis and optimum performance of two message-passing parallel processors synchronized by rollback. In E. Gelenbe, editor, *Performance '84*, pages 35–50, 1984.

[125] I. Mitrani. *Simulation Techniques for Discrete Event Systems.* Cambridge University Press, Cambridge, 1982.

[126] I. Mitrani and D. Mitra. A spectral expansion method for random walks on semi-infinite strips. In R. Beauwens and P. de Groen, editors, *Iterative Methods in Linear Algebra.* North-Holland, Amsterdam, 1992.

[127] T. Murata. Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77:541–580, 1989.

[128] R. Nelson. A performance evaluation of a general parallel processing model. *Performance Evaluation Review*, 18:13–26, 1990.

[129] R. Nelson, G. Rommel, and J.A. Stankovic. Analysis of fork-join program response times on multiprocessors. *IEEE Transactions on Parallel and Distributed Systems*, 1:286–303, 1990.

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

[131] R. Nelson and A.N. Tantawi. Approximating task response times in fork/join queues. In E. Gelenbe, editor, *High Performance Computer Systems*, pages 157–167. Elsevier Science Publishers B.V., Amsterdam, 1988.

[132] R. Nelson and D. Towsley. A performance evaluation of several priority policies for parallel processing systems. *Journal of the ACM*, 40:714–740, 1993.

[133] R. Nelson, D. Towsley, and A.N. Tantawi. Performance analysis of parallel processing systems. *IEEE Transactions on Software Engineering*, 14:532–540, 1988.

[134] 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.

[135] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models*. Johns Hopkins, Baltimore, 1981.

[136] V. Nguyen. Processing networks with parallel and sequential tasks: Heavy traffic analysis and Brownian limits. *Annals of Applied Probability*, 3:28–55, 1993.

[137] V. Nguyen. The trouble with diversity: Fork-join networks with heterogeneous customer population. *Annals of Applied Probability*, 4:1–25, 1994.

[138] K.K. Ramakrishnan. *The Design and Analysis of Resource Allocation Policies in Distributed Systems*. PhD thesis, University of Maryland, Dept. of Computer Science, 1983.

[139] M. I. Reiman and B. Simon. Open queueing networks in light traffic. *Mathematics of Operations Research*, 14:26–59, 1989.

[140] M. I. Reiman, R. J. Williams, and H. Chen. A boundary property of semimartingale reflecting Brownian motions. *Probability Theory and Related Fields*, 77:87–97, 1988.

[141] J.A.C. Resing. Polling systems and multitype branching processes. *Queueing Systems*, 13:409–426, 1993.

[142] R. Righter. Scheduling. In M. Shaked and J.G. Shanthikumar, editors, *Stochastic Orders and their Applications*. Academic Press, New York, 1994.

[143] Z. Rosberg. Deterministic routing to buffered channels. *IEEE Transactions on Communications*, 34:504–507, 1986.

[144] K.W. Ross and D.D. Yao. Optimal load balancing and scheduling in a distributed computer system. *Journal of the ACM*, 38:676–690, 1991.

[145] R.Y. Rubinstein. *Monte Carlo Optimization, Simulation and Sensitivity of Queueing Networks*. Wiley, New York, 1986.

[146] C.H. Sauer and E.A. MacNair. *Simulation of Computer Communication Systems*. Prentice-Hall, Englewood Cliffs, 1983.

[147] S. Shenker and A. Weinrib. The optimal control of heterogeneous queueing systems: A paradigm for load-sharing and routing. *IEEE Transactions on Computers*, 38:1724–1735, 1989.

[148] A. Shwartz and A. Weiss. Induced rare events: Analysis via large deviations and time reversal. *Advances in Applied Probability*, 25:667–689, 1993.

[149] P.D. Sparaggis, C.G. Cassandras, and D. Towsley. On the duality between routing and scheduling systems with finite buffer space. *IEEE Transactions on Automatic Control*, 38:1440–1446, 1993.

[150] W.J. Stewart, editor. *Numerical Solution of Markov Chains*. Marcel Dekker, New York, 1991.

[151] D. Stoyan. *Comparison Methods for Queues and Other Stochastic Models*. Wiley, New York, 1984.

[152] H. Takagi. Queueing analysis of polling models: An update. In H. Takagi, editor, *Stochastic Analysis of Computer and Communication Systems*, pages 267–318. North-Holland, Amsterdam, 1990.

[153] H. Takagi. Applications of polling models to computer networks. *Computer Networks and ISDN Systems*, 22:193–211, 1991.

[154] D. Towsley and F. Baccelli. Comparison of service disciplines in a tandem queueing network with delay dependent customer behavior. *Operations Research Letters*, 10:49–55, 1991.

[155] D. Towsley and S.S. Panwar. Optimality of the stochastic earliest deadline policy for the $G/M/c$ queue serving customers with deadlines. COINS Technical Report 91-61, University of Massachusetts at Amherst, 1991.

[156] D. Towsley and P.D. Sparaggis. Optimal routing in systems with ILR service time distributions. CMPSCI Technical Report 93-13, University of Massachusetts at Amherst, 1993.

[157] P. Tsoucas. Rare events in series of queues. *Journal of Applied Probability*, 1994. To appear.

[158] N.M. van Dijk. *Queueing Networks and Product Forms*. Wiley, New York, 1993.

[159] G.J.J.A.N. van Houtum, I.J.B.F. Adan, J. Wessels, and W.H.M. Zijm. The compensation approach for three and more dimensional random walks. Technical Report COSOR 92-39, Eindhoven University of Technology, Eindhoven, 1992.

[160] S. Varma. *Heavy and Light Traffic Approximations for Queues with Synchronization Constraints*. PhD thesis, University of Maryland, 1990.

[161] S. Varma. Performance evaluation of the time-stamp ordering algorithm in a distributed data base. *IEEE Transactions on Parallel and Distributed Systems*, 4:668–676, 1993.

[162] S. Varma and A.M. Makowski. Interpolation approximations for symmetric fork-join queues. In *Performance '93*, pages 245–273, 1993.

[163] J. Walrand. *An Introduction to Queueing Networks*. Prentice-Hall, Englewood Cliffs, 1988.

[164] Y.T. Wang and R.J.T. Morris. Load sharing in distributed systems. *IEEE Transactions on Computers*, 34:204–217, 1985.

[165] G. Weiss. Approximation results in parallel machines stochastic scheduling. *Annals of Operations Research*, 26:195–242, 1990.

[166] W. Whitt. The amount of overtaking in a network of queues. *Networks*, 14:411–426, 1984.

[167] W. Winston. Optimality of the shortest line discipline. *Journal of Applied Probability*, 14:181–189, 1977.

[168] R.W. Wolff. An upper bound for multi-channel queues. *Journal of Applied Probability*, 14:884–888, 1977.

[169] R.W. Wolff. Upper bounds on work in system for multichannel queues. *Journal of Applied Probability*, 24:547–551, 1987.

[170] P.E. Wright. Two parallel processors with coupled inputs. *Advances in Applied Probability*, 24:986–1007, 1992.

[171] S.H. Xu and H. Chen. A note on the optimal control of two interacting service stations. Working paper, 1991.

[172] U. Yechiali. Optimal dynamic control of polling systems. In J.W. Cohen and C.D. Pack, editors, *Queueing, Performance and Control in ATM*, pages 205–217. North-Holland, Amsterdam, 1991.

[173] Y. Zhao and W.K. Grassmann. The shorter queue problem: a numerical study using the matrix geometric solution. *Queueing Systems*, 8:59–79, 1991.