DE/ER/5168--1

Final Technical Report
DOE Grant DE-FG-02-93-ER25168
Approximations in the Performance Evaluation of Queueing Systems
Charles Knessl and Charles Tier, PIs
June 6, 2001

DOE Patent Clearance Granted
2/28-02
Date
Mark P. Dvorscak
(630) 252-2393
E-mail mark.dvorscak@ch.doe.gov
Office of Intellectual Property Law
DOE Chicago Operations Office

The research program on the above grant was to develop new asymptotic and perturbation methods for approximating the performance of queueing systems. This involved obtaining approximations to equations such as integral equations and difference equations. The approximations lead to highly accurate formulas for the performance measures of the systems. The models considered include buffer models (e.g. overflow and buildup), transient behavior of stochastic models, queues with time dependent parameters, diffusion approximations for queueing networks, large product-form networks, and ATM networks (Markov-modulated queues). Queueing models of these types arise in the analysis of computer and communications systems such as ATM networks which are applications in the HPCC Program. In addition, the methods developed in the proposal were also found to be applicable to other stochastic and diffusion models. Work was completed on the proposed projects as well as on new lines of research.

Below we summarize and discuss the publications which appeared during the duration of the grant.

First Passage Times, Buffer Overflows, Buildup of Large Queue Lengths: Work was completed on two of the proposed queueing problems involving first passage times and buildup of large queue lengths. The completed work is presented in the following papers.

1. C. Knessl and C. Tier, Asymptotic properties of first passage times for two tandem queues: time to empty the system, *Commun. Statist. Stochastic Models* 12, 1996, 633-672.

Two M/M/1 queues in series are considered. Assuming that at time $t = 0$ there are n_1 (n_2) customers in the first (second) queue, asymptotic expansions for the time to empty the system (residual busy period) are computed. Explicit formulas are given for the mean and variance of the emptying time, assuming that n_1 and/or n_2 are large. Recursion equations satisfied by the moments are derived and analyzed using singular perturbation techniques, such as boundary layer theory and asymptotic matching.

2. C. Knessl and C. Tier, Mean time for the development of large workloads and large queue lengths in the GI/G/1 queue, *J. Appl. Math. and Stochastic Anal.*, 9 1996, 107-142.

The GI/G/1 queue described by either the workload $U(t)$ (unfinished work) or the number of customers $N(t)$ in the system is considered. The mean time until $U(t)$ reaches or exceeds the level K is computed along with the mean time until $N(t)$ reaches N_0 . For the M/G/1 and GI/M/1 models, exact contour integral representations are obtained for these mean first passage times. The mean times are computed asymptotically, as K and $N_0 \rightarrow \infty$, by evaluating these contour integrals. For the general GI/G/1 model, asymptotic results are obtained by a singular perturbation analysis of the appropriate backward Kolmogorov equation(s). Numerical comparisons show that the asymptotic formulas are very accurate even for moderate values of K and N_0 .

Transient Behavior of Stochastic Models: Work was successfully completed on the transient behavior of certain M/M/1 queues. The completed work is contained in the following publications:

3. C. Knessl and Xiaoqian Tan, Integral representations and asymptotics for infinite and finite capacity queues described by the unfinished work I, *SIAM J. Appl. Math.* 57, 1997, 791-823.

An M/M/1 queue described by the unfinished work and also a finite-capacity version of the model is considered. Customers who cause the unfinished work to exceed a given capacity are rejected and lost. Integral representations for the time-dependent distributions are derived and evaluated asymptotically leading to a new set of approximations for the transient behavior of the system.

4. C. Knessl and Xiaoqian Tan, Integral representations and asymptotics for infinite and finite capacity queues described by the unfinished work II, *SIAM J. Appl. Math.* 57, 1997, 824-870.

This is a continuation of the study of the finite-capacity queue with customer loss.

Queues with Time-Dependent Parameters: New results were obtained for a queueing model with a time dependent arrival rate; this extends and corrects the early work of Keller and Newell and is presented in the following publication.

5. C. Knessl and Y. Yang, Asymptotic analysis of the M/G/1 queue with a time-dependent arrival rate, *Queueing Systems: Theory Appl.* 26, 1997, 23-68.

An M/G/1 queue is considered with an arrival rate that depends weakly on time, as $\lambda = \lambda(\epsilon t)$ where ϵ is a small parameter. In the asymptotic limit $\epsilon \rightarrow 0$, approximations are constructed to the probability that n customers are present at time t . The asymptotics are different for several ranges of the (slow) time scale $\tau = \epsilon t$.

Diffusion Approximations for Queueing Networks: Work was successfully completed on diffusion approximations to two queueing models. A manuscript was published that treats the diffusion approximation for two tandem queues.

6. C. Knessl and Y. Yang, Heavy traffic asymptotics of the queue length in the GI/M/1-K queue, *European J. Applied Math.*, 7, 1996, 519-543.

A GI/M/1-K queue is considered which has a capacity of K customers. Using singular perturbation methods, asymptotic approximations are constructed to the stationary queue length distribution. The capacity K is assumed to be large and several different parameter regimes are treated. Extensive numerical comparisons are used to show the quality of the proposed approximations.

7. C. Knessl and Xiaoming Tan, A fork-join queueing model: diffusion approximation, integral representations and asymptotics, *Queueing Systems: Theory and Applications*, 22, 1996, 287-322.

Two parallel M/M/1 queues are considered which are fed by a single Poisson arrival stream. An arrival splits into two parts with each part joining a different queue. The model is studied in the heavy traffic limit, where the service rate in either queue is only slightly larger than the arrival rate. The joint steady-state queue length distribution is obtained asymptotically. The symmetric and non-symmetric cases are considered and simple formulas are obtained for both.

8. C. Knessl and C. Tier, Two tandem queues with general renewal input I: diffusion approximation and integral representations, *SIAM J. Applied Math.*, 59, 1999, 1917-1959.

We consider two tandem queues with exponential servers. Arrivals to the first queue are governed by a general renewal process. If the arrivals were also exponentially distributed, this would be a simple example of a Jackson network. However, the structure of the model is much more complicated for general arrivals. We analyze the joint steady-state queue length distribution for this network, in the heavy traffic limit, where the arrival rate is only slightly less than the service rates. We formulate and solve the boundary value problem for the diffusion approximation to this model. We obtain simple integral representations for the (asymptotic) steady-state queue length distribution.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

9. C. Knessl and C. Tier, Two tandem queues with general renewal input II: asymptotic expansions for the diffusion model, *SIAM J. Applied Math.*, 59, 1999, 1960-1997.

In part I, we formulated and solved a diffusion model for two tandem queues with exponential servers and general renewal arrivals. We thus obtained the heavy traffic diffusion approximation to the steady state joint queue length distribution for this network. Here we study asymptotic and numerical properties of the diffusion approximation. In particular, analytical expressions are obtained for the tail probabilities. Both the joint distribution of the two queues and the marginal distribution of the second queue are considered. We also give numerical illustrations of how this marginal is affected by changes in the arrival and service processes.

10. C. Knessl and C. Tier, A diffusion model for two tandem queues with general renewal input, *Commun. Statist. Stochastic Models*, 15, 1999, 299-343.

We consider two tandem queues with exponential servers. Arrivals to the first queue are governed by a general renewal process. If the arrivals were also exponentially distributed, this would be a simple example of a Jackson network. However, the structure of the model is much more complicated for general arrivals. We analyze the joint steady-state queue length distribution for this network, in the heavy traffic limit, where the arrival rate is only slightly less than the service rates. We formulate and solve the boundary value problem for the diffusion approximation to this model. We obtain simple integral representations for the (asymptotic) steady-state queue length distribution. We also do a detailed study of the tail behavior of the diffusion approximation.

11. C. Knessl and C. Tier, A ray approximation to a PDE that arises in the study of tandem queues, *Studies in Applied Math.*, 102, 1999, 87-120.

We use singular perturbation methods to analyze a diffusion equation that arose in the study of two tandem queues. Denoting by $p(n_1, n_2)$ the probability that there are n_1 customers in the first queue and n_2 customers in the second queue, we obtain the approximation $p(n_1, n_2) \sim \varepsilon^2 P(X, Y) = \varepsilon^2 P(\varepsilon n_1, \varepsilon n_2)$, where ε is a small parameter. The diffusion approximation P satisfies an elliptic PDE with a non-diagonal diffusion matrix and boundary conditions that involve both normal and tangential derivatives. We analyze the boundary value problem using the ray method of geometrical optics, and other singular perturbation techniques. This yields the asymptotic behavior of $P(X, Y)$ for X and/or Y large.

12. C. Knessl, An explicit solution to a tandem queueing model, *Queueing Systems: Theory Appl.*, 30, 1998, 261-272.

Two queues in tandem, each with an exponential server and with deterministic arrivals to the first queue, are considered. An explicit solution is obtained for the steady state distribution of the process $(N_1(t), N_2(t), Y(t))$, where $N_j(t)$ is the queue length in the j th queue and $Y(t)$ measures the elapsed time since the last arrival. Then the marginal distributions of $(N_1(t), N_2(t))$ and of $N_2(t)$ are constructed. The solution is evaluated in various limiting cases, such as heavy traffic.

13. C. Knessl, A diffusion model for two parallel queues with processor sharing: transient behavior and asymptotics, *J. Appl. Math. and Stoch. Anal.*, 12, 1999, 311-338.

Two identical, parallel M/M/1 queues are considered. Both queues are fed by a Poisson arrival stream of rate λ and have service rates equal to μ . When both queues are non-empty the two systems behave independently of each other. However, when one of the queues becomes empty the corresponding server helps the server in the other queue. This is called *head-of-the-line processor sharing*. This model is

studied in the heavy traffic limit, where $\rho = \lambda/\mu \rightarrow 1$. The heavy traffic diffusion approximation is formulated and the time-dependent probability distribution of the diffusion approximation to the joint queue length process is computed. The solution is evaluated asymptotically for large values of space and/or time. This leads to simple expressions that show how the process achieves its steady state and other transient aspects.

Large Product-Form Networks: A manuscript was published that studied the asymptotic behavior of large product form networks based on the Mean Value Analysis algorithm.

14. C. Knessl and C. Tier, Asymptotic approximations and bottleneck analysis in product form queueing networks with large populations, *Performance Evaluation*, 33, 1998, 219-248.

Asymptotic approximations are constructed for the performance measures of product form queueing networks consisting of single server, fixed rate nodes with large populations. The approximations are constructed by applying singular perturbation methods to the recursion equations of Mean Value Analysis. Networks with a single job class are studied first to illustrate the use of perturbation techniques. The leading term in the approximation is related to bottleneck analysis, but fails to be accurate if there is more than one bottleneck node. A uniform approximation is constructed which is valid for networks with many bottleneck nodes. The accuracy of the uniform approximation is demonstrated for both small and large population sizes. Next, multiclass networks are considered. The leading term in the asymptotic approximation is again related to bottleneck analysis but fails to be valid across "switching surfaces". Across these the bottleneck nodes of the network change as a function of the fraction of jobs in the different job classes. A boundary layer correction is constructed near the switching surfaces which provides an asymptotic connection across the switching surfaces. Numerical examples are presented to demonstrate the accuracy of the results. We illustrate the asymptotic approach on some simple networks and indicate how to treat more complicated problems.

ATM Models: New results were obtained for several classic queueing models using the asymptotic techniques developed in our research program. We also developed approximations to important performance measures of ATM models that incorporate multiple traffic types with different service priorities. In addition, we investigated congestion control schemes such as cell discarding and traffic shaping.

The new results are presented in:

15. C. Knessl and C. Tier, Heavy traffic analysis of a Markov-modulated queue with finite capacity and general service times, *SIAM J. Appl. Math.*, 58, 1998, 257-323.

An ATM model with a set of N independent sources, each of which alternates between "on" and "off" states is considered. When a source is on it generates a Poisson arrival stream to a finite capacity queue with a general server. The balance equations satisfied by the joint steady-state distribution of the queue length and the number of on sources are derived. The problem is analyzed in the heavy traffic limit where $N \rightarrow \infty$. Numerical and asymptotic methods are developed for solving the PDEs. The analysis makes use of singular perturbation techniques and special functions.

16. D.I. Choi, C. Knessl and C. Tier, A queueing System with queue length dependent service times, with applications to cell discarding in ATM networks, *J. Appl. Math. Stochastic Anal.*, 12, 1999, 35-62.

A queueing system $(M/G_1, G_2/1/K)$ is considered in which the service time of a customer entering service depends on whether the queue length, $N(t)$, is above or below a threshold L . The arrival process is Poisson and the general service times S_1 and S_2 depend on whether the queue length at the time service is initiated is $< L$ or $\geq L$, respectively. Balance equations are given for the stationary

probabilities of the Markov process $(N(t), X(t))$ where $X(t)$ is the remaining service time of the customer currently in service. Exact solutions for the stationary probabilities are constructed for both infinite and finite capacity systems. Asymptotic approximations of the solutions are given, which yield simple formulas for performance measures such as loss rates and tail probabilities. The numerical accuracy of the asymptotic results is tested.

17. C. Knessl and C. Tier, A Simple Fluid Model for Servicing Priority Traffic, to appear *IEEE Transactions on Automatic Control*, 2001.

A simple priority queueing system is considered, with two different types of traffic, high and low priority. Each type of traffic generates arrivals to its own buffer that are modeled by continuous fluid flows. The input flow into the high priority buffer is Markov modulated while the input rate into the low priority buffer is constant. The fluid is transmitted from the buffers with the high priority buffer having priority over the low priority buffer. The problem for the joint distribution of the buffer contents is derived and solved explicitly. Approximations are constructed for the tail behavior of the buffer content.

Related Stochastic and other Models: Asymptotic methods of the type developed for queueing models have also been successfully applied to models in other areas. The broad range of applications indicates the power of our asymptotic approach.

18. C. Knessl and C. Peters, Exact and asymptotic solutions for the time dependent problem of collective ruin II, *SIAM J. Appl. Math.* 56, 1996, 1471-1521.

A model for the risk reserve of an insurance company is considered. It is assumed that the reserve increases due to premiums and also earns interest. The reserve decreases due to claims which are modeled as a compound Poisson process. The asymptotic properties of the probability that the reserve remains positive are studied.

19. J.B. Keller and C. Knessl, Advection-diffusion past a strip. I. normal incidence, *J. Math. Phys.* 38, 1997, 267-282.

The concentration of particles moving by diffusion and advection or drift is analyzed. The motion is impeded by an impenetrable strip which is parallel to the z -axis. It is assumed that the concentration of particles satisfies a linear advection-diffusion equation with boundary conditions on the strip. This problem is solved for large values of vL/d , where v is the drift velocity, D is the diffusion coefficient, and $2L$ is the width of the strip. Methods of asymptotic analysis are used.

20. J.B. Keller and C. Knessl, Advection-diffusion past a strip. II. oblique incidence, *J. Math. Phys.* 38, 1997, 902-925.

Advection and diffusion of particles past an impenetrable strip is considered when the strip is oblique to the advection or drift velocity. The particle concentration is determined asymptotically for large values of vL/d , where v is the drift velocity, D is the diffusion coefficient, and $2L$ is the width of the strip.

21. C. Knessl, A note on the asymptotic behavior of the depth of tries, *Algorithmica*, 22, 1998, 547-560.

The depth distribution of digital trees called tries is analyzed. Assuming that the trie is constructed from n statistically independent strings, the probability that the depth is equal to k is studied asymptotically for n and/or k large. Detailed results are obtained for $n \rightarrow \infty$ and various ranges of k . An accurate description of the tails of the probability distribution is given. If the symbols in the string are 0's and 1's and they occur with respective probabilities q and $p = 1 - q$, the symmetric, non-symmetric and "nearly symmetric" models are studied. For the latter model, a new limiting distribution is obtained.

22. D.S. Ahluwalia, J.B. Keller and C. Knessl, Advection-diffusion around a curved obstacle, *J. Math. Phys.* 39, 1998, 3694-3710.

Advection and diffusion of a substance around a curved obstacle is analyzed when the advection velocity is large compared to the diffusion velocity, i.e. when the Peclet number is large. Asymptotic expressions for the concentration are obtained by the use of boundary layer theory, matched asymptotic expansions, etc. The results supplement and extend previous ones for straight obstacles. The apply to electrophoresis, the flow of ground water, chromatography, sedimentation, etc.

23. C. Knessl, Integral representations and asymptotic expansions for Shannon and Renyi entropies, *Appl. Math. Lett.*, 11, 1998, 67-74.

Integral representations for the Shannon and Renyi entropies associated with simple probability distributions are derived. These include Poisson, binomial, and negative binomial distributions. Then full asymptotic expansions for the entropies are obtained.

24. C. Knessl and W. Szpankowski, Quicksort algorithm again revisited, *Discrete Math. and Theoretical Computer Science*, 3, 1999, 46-63.

The standard Quicksort algorithm is considered that sorts n distinct keys with all possible $n!$ orderings of keys being equally likely. Equivalently, the total path length $\mathcal{L}_n$ in a randomly built *binary search tree* is analyzed. Obtaining the limiting distribution of $\mathcal{L}_n$ is still an outstanding open problem. In this paper, an integral equation is derived for the probability density of the number of comparisons $\mathcal{L}_n$. Then, the large deviations of $\mathcal{L}_n$ are investigated. The left tail of the limiting distribution is shown to be much "thinner" (i.e. double exponential) than the right tail (which is only exponential). The results contain some constants that must be determined numerically. Formal asymptotic methods of applied mathematics such as the WKB method and matched asymptotics are used.