Iterative approximation for networks of queues

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics and Computing Science

Memorandum COSOR 83 - 22

Iterative approximations for networks of queues

by

J. van Doremalen

J. Wessels

Eindhoven, the Netherlands

November 1983

ITERATIVE APPROXIMATIONS FOR NETWORKS OF QUEUES

Jan van Doremalen and Jaap Wessels
Eindhoven, 1983

Abstract. If networks of queues satisfy certain conditions, then the equilibrium distribution for the number of jobs in the various stations has the so-called product-form. In such cases there are relatively elegant and simple computational procedures for the relevant behavioral characteristics. Quite commonly, however, the conditions are too severe and exact solution is practically impossible for larger problems.
In this paper we will consider iterative approximations for networks of queues which either don't possess product-form solutions or are so large that exact solution becomes intractable even using the product-form of the solution. The approximations are based on a mean value analysis approach and use either aggregation of some sort or decomposition. For the details of the approximations heuristic arguments are used. The approach is worked out for some problem types.

1. Introduction

In many areas networks of queues are used as models: production planning in manufacturing enterprises, computer performance evaluation, design of communication networks, planning of harbour facilities, etc. General queueing theory does not provide much help for the analysis of such complex queueing models. The only help can be found in the line of research that emerged from Jackson's paper [10] in which it was proved that the equilibrium distribution for a particular type of networks has a product-form. Extending Jackson's result it has been proved that a large class of networks has equilibrium probabilities with a product-form (confer Kelly [11]). It has also been shown that for such queueing networks the relevant behavioral characteristics can be computed in some (relatively) simple and elegant ways. The two main procedures are known by the name of convolution method (confer Reiser and Kobayashi [16]) and by the name of mean value analysis (confer Reiser and Lavenberg [17] and Reiser [15]).

Regrettably, however, many practical problems do not satisfy the conditions for having product-form solutions, whereas other problems are very large and therefore intractable using the standard methods. For both types of problems the only way out seems to be approximation. Several methods of approximation have been published. For instance approximate decomposition (confer Courtois [4]) which is used in the handling of memory queues in computer evaluation studies (confer Hine, Mitrani and Tsur [9])

and for handling FIFO-servers with arbitrary service time distributions in open networks (confer Kühn [12]). For an overview of several approaches see Chandy and Sauer [3].

In recent years the mean value analysis procedure has become popular as a basis for approximation. For a recent overview and appraisal see de Souza a Silva, Lavenberg and Muntz [19]. Although the approximation methods for different types of problems show some structural resemblance, the methods are basically heuristic. Only in some cases one has been successful in obtaining convergence and uniqueness results (see [19] for some examples and further references).
In this paper we will present heuristics and numerical results for two types of problems and discuss the same topics for some other problems.
The first problem, which will be treated in Section 2, is a rather specific one. It arised in treating the planning of harbour facilities, where it appeared to be necessary to include servers with a two-phase service procedure. The first phase is a preparatory one and may be executed for the first customer of a busy period in the preceding idle period. This feature destroys the product-form.
The next problem has attracted a lot of attention in the literature: the problem of many customer chains in a closed network. The conditions for the existence of a product-form solution are not violated, but even the efficient mean value analysis procedure requires too much work if the number of chains is relatively large. To obtain approximations, the usual approach is to remove the recursion from the mean value scheme. In Section 3, we will present a decomposition approach, which maintains the recursion, but transforms the multidimensional recursion in several one-dimensional recursions.
For both problems numerical results are compared to exact solutions. For the second problem a comparison with other methods will be given also.
In Section 4 some experience with other methods will be reported. Here, as well as in Section 2, the heuristics are basically some sort of aggregation. Disaggregation provides the basis for the next iteration step.

2. The two-phase server with preparatory first phase

Consider a closed queueing network with N single server FIFO stations in which K customers walk around with routing probabilities p_{mn} for going from station m to station n. At station n the customers have negative exponentially distributed workloads with expected value w_n. The network satisfies the conditions for having a product-form solution. The arrival theorem for such networks says that a customer sees the system upon a jumpmoment as if it is the system with K - 1 customers in equilibrium. We can use this theorem to express the mean residence time $S_n(K)$ at

queue n in the mean number of customers $L_n(K-1)$ at that queue if there are $K-1$ customers in the system,

$$(1) \qquad S_n(K) = L_n(K-1)w_n + w_n \ .$$

The RHS denotes the average amount of work a customer sees in front of him upon his arrival at queue n plus his own work. Applying Little's formula to queue n, we have

$$(2) \qquad L_n(K) = \Lambda_n(K)\, S_n(K) \ ,$$

where $\Lambda_n(K)$ denotes the throughput at queue n. Finally, applying Little's formula over the network,

$$(3) \qquad \Lambda_n(K) = \vartheta_n K \Big[\sum_{m=1}^{N} \vartheta_m S_m(K) \Big]^{-1} \ ,$$

where the ϑ_n's are the unique solution of

$$(4) \qquad \vartheta_n = \sum_{m=1}^{N} \vartheta_m p_{mn} \qquad \text{and} \qquad \sum_{n=1}^{N} \vartheta_n = 1 \ .$$

Starting with $L_n(0) = 0$ these relations give a recursive scheme to evaluate the mean values. For more details on this mean value scheme and the arrival theorem we refer to Reiser and Lavenberg [17] and Reiser [15]. If we introduce an extraordinary behaviour at one of the stations, for example non-exponential service times, formulae (2), (3) and (4) remain valid. However, relation (1) will be violated. To some extent the idea behind the relation will remain and, therefore, it seems sensible to consider a mean value scheme with a slightly adjusted form of relation (1) to incorporate the effects of the extraordinary behaviour.

As an example of such a deviant behaviour we will consider a network where some server n may have a workload, which per customer consists of two negative exponentially distributed phases, $w_n = w_n' + w_n''$. The first phase is a kind of preparatory one and can be started (and sometimes be completed) during an idle period. Thus the first customer of a busy period has a different workload and the effect will be that some of the customers only experience a workload w_n'', whereas others have the full workload $w_n' + w_n''$.

The steady-state probabilities no longer have a product-form, but the network still can be analyzed as a continuous-time Markov-process on a finite state space. To

solve for the corresponding set of equilibrium equations is very unattractive from a computational point of view. We will develop an iterative approximation based on the mean value scheme and an adjustment of relation (1).

The first guess in adapting Formula (1) seems to be to maintain $L_n(K-1)$ as the expected number of customers present upon arrival (this need not be true) and to replace w_n by an adjusted value,

$$\tilde{w}_n = (1 - a_n)w_n' + w_n'' , \tag{5}$$

where a_n denotes the probability that an arriving customer finds his preparatory phase already completed. Thus we implicitly assume that all customers have the same negative exponentially distributed workload with mean $\tilde{w}_n$, i.e. we approximate the original model by a model with a product-form solution. To find a_n requires a rigorous analysis of the original problem and that we just wanted to avoid. However, one may make a guess, for instance $a_n = 0$ or $a_n = 1$, and try to improve the guess after an evaluation of the mean value scheme. Suppose we have an initial guess for a_n and we have solved the mean value scheme (1) through (4) with w_n replaced by $\tilde{w}_n$. How to improve on the initial guess for a_n? The true a_n can be written as

$$a_n = b_n c_n \tag{6}$$

with b_n the probability that an arriving customer is the first one in a busy period and c_n the probability that a preparatory phase is completed before the end of an idle period. Better estimates for b_n and c_n then can be constructed as follows (confer van Doremalen and Wessels [7]),

$$b_n' = 1 - \Lambda_n(K-1)\tilde{w}_n \tag{7}$$

$$c_n' = w_n'(w_n' + v_n')^{-1} \tag{8}$$

where

$$v_n' = (1 - \Lambda_n(K)\tilde{w}_n)(\Lambda_n(K)b_n')^{-1} . \tag{9}$$

The results of the iteration scheme are fairly good, particularly for the throughput. Mean queue lengths and mean residence times in general are less accurately approximated. As a simple numerical example the results of a cyclical network with three stations are depicted in Table 1. Evaluated are the exact throughput, a lowerbound ($a_1 = 1$), an upperbound ($a_1 = 0$) and the approximation resulting from the iterative method. The last column gives the limiting values for a_1.

w_2	w_3	throughputs				a_1
		exact	low	appr.	high	
2	2	.326	.300	.321	.347	.42
8	8	.093	.091	.092	.093	.83
.25	.25	.500	.497	.499	.946	.01

Table 1. Throughputs in a cyclical network with one two-phase server $w_1' = w_1'' = 1$, $K = 3$ and $N = 3$.

It is possible to refine these results. One way would be to use Kühns decomposition approach, confer Kühn [12], to take into account the non-exponential character of the two phase servers. A natural extension of the method then is to consider the case that the phases themselves are non-exponential.

3. Closed multichain queueing networks

Again consider a closed network with N single server FIFO stations. Now there are R irreducible customer chains, where the K_r customers of chain r have routing probabilities p_{mn}^r for going from station m to station n. At station n all customers have negative exponentially distributed workloads with the same expected value w_n. The arrival theorem states that a customer sees upon a junpmoment the system in equilibrium as if one customer of his own chain has been removed. If we denote the population-vector $(K_1,\ldots,K_r)$ as K, this theorem implies that $S_{nr}(K)$, the mean residence time of a chain r customer at station n, can be expressed in $L_{n\ell}(K - e_r)$, the mean number of chain ℓ customers at station n if one customer of chain r has been removed from the system,

$$S_{nr}(K) = \sum_{\ell=1}^{R} L_{n\ell}(K - e_r) w_n + w_n \quad . \tag{10}$$

Application of Little's formula to station n gives,

$$L_{nr}(K) = \Lambda_{nr}(K)\, S_{nr}(K) \quad , \tag{11}$$

where $\Lambda_{nr}(K)$ denotes the throughput of chain r customers at queue n. Finally, the multichain-equivalent of Relation (3) is,

$$(12) \qquad \Lambda_{nr}(K) = \vartheta_{nr} K_r \Big(\sum_{m=1}^{N} \vartheta_{mr} S_{mr}(K) \Big)^{-1}$$

where the ϑ_{nr}'s are, for $r = 1,2,\ldots,R$, the unique solution of

$$(13) \qquad \vartheta_{nr} = \sum_{m=1}^{N} \vartheta_{mr} p_{mn}^{r} \qquad \text{and} \qquad \sum_{n=1}^{N} \vartheta_{nr} = 1 \ .$$

For more details on the multichain mean value scheme we refer to Reiser and Lavenberg [17].

The recursion, defined by the Relations (10) through (13), now runs through all vectors in the range from $(0,\ldots,0)$ to $(K_1,\ldots,K_R)$. The storage requirements and the complexity of the algorithm grow exponentially with the number of chains. The apparent problem differs essentially from the one described in Section 2. Now the product-form solution is not violated, but the complexity of the algorithm prohibits an exact evaluation for larger values of $R,K_1,\ldots,K_R$ and approximate methods have to be recommended for that reason.

In the literature several approximation methods have been considered, e.g. by Schweitzer [18], Reiser [14], Reiser and Lavenberg [17] and Chandy and Neuse [2]. Very recently, an overview of these and other methods appeared in de Souza a Silva, Lavenberg and Muntz [19]. The usual approach is to remove the recursion from the mean value scheme and to concentrate on an iterative approximation of the mean values at the population vector K. We will exploit a decomposition idea in which R single chain networks are analyzed. Iteratively an improved approximation of the mutual influence of the chains is incorporated in the single chain analysis.

For chain r, $r = 1,2,\ldots,R$, consider the following adjusted single chain mean value scheme. Evaluate for $k = 1,2,\ldots,K_r$,

$$(14) \qquad S_{nr}^{*}(k) = L_{nr}^{*}(k-1)w_n + w_n + A_{nr}(k)w_n$$

$$(15) \qquad \Lambda_{nr}^{*}(k) = \vartheta_{nr} k \Big(\sum_{m=1}^{N} S_{mr}^{*}(k) \Big)^{-1}$$

$$(16) \qquad L_{nr}^{*}(k) = \Lambda_{nr}^{*}(k) \, S_{nr}^{*}(k)$$

where the factor $A_{nr}(k)$ denotes the number of customers of other chains a chain r

customer sees in front of him upon arrival at station n if k customers of his own chain are in the system. As an approximation for $A_{nr}(k)$ we propose

$$(17) \qquad A_{nr}(k) = \sum_{\ell \neq r} L^{*}_{n\ell}(K_{\ell})$$

where we use as an approximation assumption that a chain r customer sees the other chains as if in global equilibrium. Equations (14) through (17) implicitly give the approximations for the mean values. A standard technique to solve for these equations is to start with initial values for the $A_{nr}(k)$'s and to iterate the scheme for the successive chains until convergence is established. Using Brouwer's Fixed Point Theorem one can prove the existence of a positive solution of the equations. Up till now we have not been able to prove uniqueness of the solution and convergence of the method. However, numerical experiments show a relatively fast convergence and the approximations usually are within a few percent of the exact values. One can construct examples where the approximations are rather poor.
We will show a numerical example where we have compared the exact results of the mean value scheme with four different approximation methods.

Consider the model of a computer system with three terminal groups pictured in Figure 1. The system consists of a central processor unit (CPU) and three disk-groups (D1, D2 and D3). The service discipline at these four stations is first-in first-out and the exponential workloads have expected values 10 m sec, 20 m sec, 20 m sec and 30 m sec respectively. There are three terminal groups (T1, T2 and T3). The 20 active terminals of T1 have mean think times of 10 sec. They generate requests which in the average have 20 CPU calls, 15 D1 calls and 4 D2 calls. A terminal starts thinking again if his request has been handled and a response has been returned. The 10 active terminals of T2 have thinktimes of 20 sec, and requests of 40 CPU calls, 14 D1 calls abd 25 D2 calls. The 10 active terminals of T3 have thinktimes of 60 sec and requests of 200 CPU calls, 20 D1 calls, 40 D2 calls and 139 D3 calls.

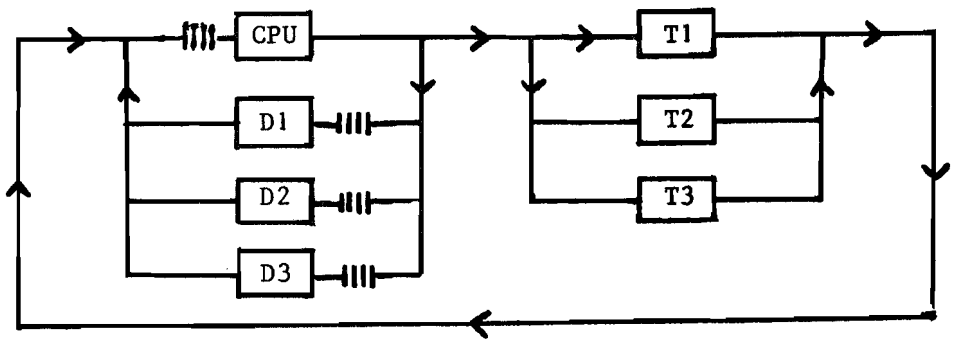

Figure 1. A computer system with terminal groups.

In Table 2 are pictured the utilizations of the CPU and the disk groups. The exact results are compared with four approximation methods, namely the methods of Schweitzer (SCHW), Reiser (R), Reiser and Lavenberg (R+L) and our method (D+W). In Table 3 are the response times for the three terminal groups. We note that the four methods all perform very good. At the moment we are studying other applications and examples and a more detailed report on the methods is in preparation. We remark that our method very straightforwardly can be extended to systems with LIFO (last-in first-out), PS (processor-sharing) and IS (infinite server) stations. This, for example, makes it possible to consider the above model with a processor-sharing CPU and consequently with different workloads at the CPU for the different terminalgroups. Finally, it should be noted that the method can be extended to mixed open and closed networks, confer van Doremalen [6].

	exact	SCHW	R	R+L	D+W
CPU	.774	.766	.770	.774	.768
D1	.686	.679	.683	.687	.682
D2	.457	.452	.454	.457	.454
D3	.539	.532	.536	.537	.535

Table 2. Utilizations in the computer system.

	exact	SCHW	R	R+L	D+W
T1	1.69	1.81	1.74	1.67	1.77
T2	3.11	3.29	3.22	3.02	3.26
T3	17.30	18.44	17.79	17.67	17.95

Table 3. Response times of the terminal groups.

4. Some other applications

In this section we will venture on some other applications of approximation techniques involving blocking phenomena, priority rules and FIFO stations with class dependent workloads.

4.1. Blocking

Consider the network model of Section 2. Now at queue n only a restricted number b_n of customers is allowed for. The joining of queue n is forbidden as long as b_n customers are present. A customer not allowed to enter station n, waits in the originating server and blocks this server until the unblocking moment. The effect of blocking is a decrease in the availability of the blocked servers. This can be accounted for by increasing the workloads at the blocked servers with some factor which may be determined iteratively, using estimates for the blocking probability from the preceding analysis. The results obtained so far, show an improvement compared to the total neglection of blocking effects. Especially, if the effect of blocking is not too heavy the approach seems to work quite well. A detailed report on this case is in preparation.

4.2. Priorities

Consider the model of Section 3. However, now there is some kind of priority for certain chains at certain queues. We thereby can think of preemptive-resume priorities and head-of-the-line priorities. Non-iterative approximations for such models for example can be based on the mean value analysis of M/G/1-priority queues as described in van Doremalen [5]. Iterative approximations might be based on the convolution algorithm. Results obtained so far are very promising and research in that direction is in progress.

4.3. Chain dependent workloads at FIFO single server stations

The mean value scheme of Section 3 for closed multichain queueing networks works alright if we assume the same negative exponentially distributed service times for all customer chains at a specific station. However, if the mean service times w_{nr} for the chains at a certain queue n do differ, the product-form solution no longer holds and the mean value scheme does not give exact results.

One way out is a relaxation of the mean value scheme. This straightforward non-iterative approximation has been considered by others also, confer Bard [1]. Instead of Relation (10) we get

$$(18) \qquad S_{nr}(K) = \sum_{\ell=1}^{R} L_{n\ell}(K - e_r) w_{n\ell} + w_{nr} \ .$$

Another method is the well-known processor-sharing approximation which reduces to the following, intuitively less attracting, adjustment of (10),

$$(19) \qquad S_{nr}(K) = \sum_{\ell=1}^{R} L_{n\ell}(K - e_r)w_{nr} + w_{nr} .$$

Numerical experiments show the first method to be considerably better. A totally different approach is to use a negative exponential service time distribution with a mean which is a proper mixture of the original means. Iteratively, this mixture can be determined. The results are not too well and it seems better to use explicit estimates for the probability that the server works on a particular type of job. A report on such an approach is in preparation.

5. Concluding remarks

We have considered the use of iterative approximation methods in several applications The importance of approximation methods in the analysis of queueing networks is paramount for several reasons.

First of all, exact analysis is limited to only a few restricted models as for example the networks which satisfy the conditions for the existence of a product-form solution for the steady-state probabilities. Though this class of networks still is subject of research and techniques are being developed to extend the class (confer Kelly [11] and van Dijk and Hordijk [8]), it is clear that very important classes of networks never will be fitted in this frame.

But, as we have seen in Section 3, there is another problem. Even for models in a class which can be analyzed elegantly, the amount of work to be done can prohibit an exact evaluation of important performance measures. Of course, one can try to improve the evaluation methods as for instance has been done by Lam and Lien [13], but again there always will be the need of fast approximation methods.

References

[1] Y. Bard, Some extensions to multichain queueing network analysis. 4th Int. Symp. on Modelling and Performance Evaluation of Computer Systems, Vienna 1979.

[2] K.M. Chandy and D. Neuse, Lineariser: A heuristic algorithm for queueing network models of computing systems. Comm. of the A.C.M. 25 (1982) 126 - 134.

[3] K.M. Chandy and C.H. Sauer, Approximate methods for analyzing queueing network models of computing systems.
Computing Surveys 10 (1978) 281 - 317.

[4] P.J. Courtois, Decomposability: Queueing and Computer System Applications.
Academic Press, New York 1977.

[5] J. van Doremalen, A mean value approach for M/G/1 priority queues.
Memorandum COSOR 83-09, Eindhoven University of Technology 1983.

[6] J. van Doremalen, Mean value analysis in multichain queueing networks: an iterative approximation.
DGOR Operations Research Proceedings 1983, Springer Verlag, Berlin.
To appear.

[7] J. van Doremalen and J. Wessels, An iterative approximation for closed queueing networks with two-phase servers.
Memorandum COSOR 83-12, Eindhoven University of Technology 1983.

[8] N. van Dijk and A. Hordijk, Networks of queues: Part I, Job-local-balance and the adjoint process. Part II, General routing and service characteristics.
Proc. of the Int. Sem. on Modelling and Performance Evaluation Methodology, Paris 1982.

[9] J.H. Hine, I. Mitrani and S. Tsur, The control of response times in multiclass systems by memory allocation.
Comm. of the A.C.M. 22 (1979) 415 - 424.

[10] J.R. Jackson, Networks of waiting lines.
O.R. 5 (1957) 518 - 521.

[11] F.P. Kelly, Reversibility and stochastic networks.
John Wiley and Sons, New York 1978.

[12] P.J. Kühn, Approximate analysis of general queueing networks by decomposition.
IEEE Trans. Comm. 27 (1979) 113 - 126.

[13] S. Lam and Y. Lien, A tree convolution algorithm for the solution of queueing networks.
Comm. of the A.C.M. 26 (1983) 203 - 215.

[14] M. Reiser, Mean value analysis: A new look at an old problem.
4th Int. Symp. on Modelling and Performance Evaluation of Computer Systems, Vienna 1979.

[15] M. Reiser, Mean value analysis and convolution method for queue-dependent servers in closed queueing networks.
Performance Evaluation 1 (1981) 7 - 18.

[16] M. Reiser and H. Kobayashi, Queueing networks with multiple closed chains: theory and computational algorithms.
IBM J. Res. Dev. 19 (1975) 283 - 294.

[17] M. Reiser ans S.S. Lavenberg, Mean value analysis of closed multichain queueing networks.
Comm. of the A.C.M. 27 (1980) 313 - 322.

[18] P. Schweitzer, Approximate analysis of multiclass networks of queues. Presented at the Int. Conf. on Stochastic Control and Optimization, Amsterdam 1979.

[19] E. de Souza a Silva, S.S. Lavenberg and R.R. Muntz, A perspective on iterative methods for the approximate analysis of closed queueing networks. Proc. Int. Workshop on Applied Mathematics and Performance Reliability Models of Computer Communication Systems, University of Pisa 1983.