Queueing Systems 43, 307–328, 2003

On Markovian Multi-Class, Multi-Server Queueing

A. VAN HARTEN * and A. SLEPTCHENKO a.vanharten@sms.utwente.nl
University of Twente, Faculty of Technology and Management, P.O. Box 217, 7500 AE Enschede, The Netherlands

Received 7 October 2000; Revised 5 September 2002

Abstract. Multi-class multi-server queueing problems are a generalisation of the well-known $M/M/k$ queue to arrival processes with clients of N types that require exponentially distributed service with different average service times. In this paper, we give a procedure to construct exact solutions of the stationary state equations using the special structure of these equations. Essential in this procedure is the reduction of a part of the problem to a backward second order difference equation with constant coefficients. It follows that the exact solution can be found by eigenmode decomposition. In general eigenmodes do not have a simple product structure as one might expect intuitively. Further, using the exact solution, all kinds of interesting performance measures can be computed and compared with heuristic approximations (insofar available in the literature). We provide some new approximations based on special multiplicative eigenmodes, including the dominant mode in the heavy traffic limit. We illustrate our methods with numerical results. It turns out that our approximation method is better for higher moments than some other approximations known in the literature. Moreover, we demonstrate that our theory is useful to applications where correlation between items plays a role, such as spare parts management.

Keywords: queueing, several types of clients, performance analysis, Markov chains, steady state analysis

1. Introduction

Multi-class, multi-server (MCMS) models arise when clients with different service characteristics ask for the same capacity and performance characteristics are needed for each class separately. Applications arise, e.g., in manufacturing, when a workstation has to process different job types with each their own work content and one is interested in the throughput time per job type, cf. [3,10,12]. We encountered MCMS queues in our research on inventory control of repairable spare parts when modeling repair facilities [15,16]. Particularly, we needed the first two moments of the number of items in the system, or the backorders, per class and the correlation between classes. These quantities have a direct relation to the availability of the installed base served by the repair facility.

Although some literature on MCMS queues is available, there are limitations on the number of classes and/or the number of servers and on the performance characteristics considered (see the next section for a literature review). Here we present an exact algorithm to compute the steady state probabilities in Markovian systems as de-

* Corresponding author.

scribed, i.e. each customer class has its own Poisson arrival process and arrival rate as well as its own exponentially distributed service time. From these state probabilities, we can derive a large variety of relevant class-dependent performance characteristics. Of course, it is rather straightforward to write down the global balance equations of the corresponding Markov chain. However, it is not trivial to find the exact solution of these infinite-dimensional equations. New in our approach is that we show that the exact solution has a *special structure.* This allows us to reduce the construction of the solution to finding the eigenvalues and eigenvectors of a *finite*-dimensional matrix, where in a Wiener–Hopf-like way only half of the eigenvalues corresponding with decay for increasing numbers of clients are relevant. In this paper, we explain our approach in detail and we show the results in some numerical experiments. Also, we compare our exact results to some approximations that have been described in the literature. The exact computations for high numbers of servers and/or classes are lengthy and good, numerically more efficient approximations are useful. We define some approximations based on a restricted set of multiplicative eigenmodes. In this way we can demonstrate a new sort of multiplicative state space collapse in the heavy traffic approximation for non-dedicated multi-server systems (see next section for a literature survey). Moreover, our approximation has advantages for higher moments compared with [7,17].

The remainder of this paper is structured as follows. First, we give a brief overview of the relevant literature and we explain our contribution in section 2. Next, we sketch our approach and we derive the steady state equations (section 3). Then, we solve these equations (including the nontrivial eigenvalues and eigenmodes) explicitly for the special cases of equal service rates (section 4). We discuss the exact solution of the general model in section 5. Next, we define our approximate, restricted mode solution in section 6. In section 7, we derive relevant performance measures from the steady state probabilities. Numerical results and computational efficiency are discussed in section 8. In this section we also discuss an application of our results to obtain better approximations for the system availability of an installed base taking correlation between various items into account. Moreover we discuss the quality of our "multiplicative modes only" approximation compared with others. Finally, we give our conclusions in section 9.

2. Literature

There are only few papers that are directly devoted to MCMS models, as far as we know [5,6]. De Smit [5] analyses the $GI/M/k$ multi-class queue where different types of customers have different service rates. Here GI refers to general arrival processes and H_{m} denotes a mixture of different exponential distributions, as one obtains in a multi-class model, if one abstracts from the identity of the different jobs and introduces one overall jobtype. His approach uses phase vectors and a solution is derived by Laplace transform and Wiener–Hopf decomposition techniques. He obtains explicit results for the stationary distributions of waiting times and the queue length. It is shown that the stationary waiting time distribution is a mixture of exponentials, generalizing a previous result for $M/H_{\mathrm{m}}/2$ of Cohen [4]. This result is theoretical, but de Smit [6] develops a

numerical solution based on these results. However, the numerical procedure is limited to H_2 service time distributions or equivalently a two-class system with exponential service time distributions.

Another line of research on the MCMS queue is the application of approximation techniques or asymptotic methods. Bertsimas and Mourtzinou [1] consider systems with general arrival and service distributions in heavy traffic. However, their results are restricted to single server systems. Diaz and Fu [7], give some approximations mainly for the single server case. Basic to the reasoning underlying their approximations is that the expected waiting time is class-independent, see also [17]. Moreover, they assume the correlation between items in queue and in service is related in a simple way to the class utilisation fractions. As we will see in the remainder of this thesis, the exact solution of MCMS problems has a more complex structure. Other interesting work in this respect is (Adan and van der Wai, 1998). They relate the mean lead time of a two-class system to those of a single class system with a Coxian-2 service process. They show that systems where different items are processed by the same servers might perform better than the system with dedicated servers if the difference between service rates (μ_1/μ_2) is not too big (>7).

Recently quite some work has been done on heavy traffic approximations for multi-class queueing systems with dedicated servers using fluid model equations and semi-martingales. In [2,18] a proof of a state space collapse is given under certain, rather general conditions. They show that that these conditions are satisfied for FIFO networks of Kelly type (class independent service time distributions). Our approximate solution based on multiplicative eigenmodes gives rise to an analogous effect for non-dedicated MCMS systems. It turns out that the critical eigenmode in case of heavy traffic is of multiplicative type and the approximation error vanishes for the utilisation approaching 1. State space collapse occurs in the sense that the critical eigenmode dominates the queueing behaviour.

The contributions of our paper to the existing literature are the following:

1. We use a direct approach to solving the steady state equations exactly without restriction on the number of classes or the number of servers (although the combination of many servers and many customer classes may require long computer run times).

2. We are able to derive a wide range of performance measures, not only the standard measures as the mean waiting time or the mean queue length. Particularly, we can easily calculate higher moments of the queue length and the number of items (or backorders) in the system per class. We can even derive correlation coefficients, e.g., between the number of items (or backorders) in the system for different classes.

3. We show that the exact solution of the steady state equations has a special structure, which up to the best of our knowledge has not been recognised before in the existing literature. Particularly, we show that the solution does not have a simple product structure as one might expect intuitively.

4. We present a new sort of approximation based on a restricted number of multiplicative eigenmodes, which contains the state space collapse behaviour in the heavy traffic limit and we discuss its merits.

3. Stationary state equations for MCMS problems

Let us now describe the problem in more detail. Jobs with type i arrive according to a Poisson process with arrival rate λ_i. The total arrival rate is given by $\Lambda = \sum \lambda_i$. The arrival fraction of class i is $a_i = \lambda_i/\Lambda$. We consider identical, non-dedicated servers, hence the flow through a server consists of all job types. The service distribution of job type i is exponential with rate μ_i, and it is the same for all servers. The service discipline by which jobs are assigned is first-come-first-serve (FCFS). Hence the stability condition for this queueing system is given by $\rho < 1$. The average service rate μ is defined by $1/\mu = \sum a_i/\mu_i$. The utilisation is represented by ρ with $\mu = \Lambda/(k\rho)$. Our notation for the relative perturbation from the average service rate is δ_i with $\mu_i = \mu(1 + \delta_i)$.

To describe the state of the queueing system, we will use two vectors $\overline{w}$ and $\bar{s}$ of dimension N, i.e. each component i contains information about the amount of items of each class i in queue or in service, correspondently. Because of the PASTA property (Poisson Arrivals See Time Averages), these vectors will provide us all the information about the system that we need. Let us now have a closer look at the stationary state equations of the MCMS system.

The state equations follow from a micro-balance reasoning as usual. The net exchange of probability in an infinitesimal time interval from a given state with its neighbours has to be zero in an equilibrium situation. Neighbours of a state $(\overline{w}, \bar{s})$ with n clients are states to or from which a one step transition is possible, either by an arrival event (A) or a service completion event (C). Neighbour states have $n-1$ or $n+1$ clients.

Before stating the stationary state equations, we first introduce the following notation. We use $|\overline{w}|$ for the total number of items in the queue, analogously we use $|\bar{s}|$ for the total number of items in service. The next expressions are obvious:

$$|\overline{w}| + |\bar{s}| = n; \qquad |\overline{w}| = 0, \quad \text{if } n \leqslant k; \qquad |\bar{s}| = k, \quad \text{if } n \geqslant k.$$

The vector $\bar{e}_i$ is defined with all zero components except for the component i, which is 1. In addition to this vector, we define e_{ij} as 1 if $i = j$ and 0 otherwise. Also, we define

$$\delta(\bar{s}) \stackrel{\text{def}}{=} \frac{1}{k} \sum_{i=1}^{N} s_i \delta_i.$$

Using the structure of the arrival process, the stationary equations for $n > k$ can be given in a reduced form. For $n - k$ consecutive arrivals, the probability distribution over the possible vectors describing the queue $\overline{w}$ is given by the product of the arrival

fractions and by the number of possible arrival sequences within the vector $\overline{w}$. As a consequence, we may write the steady state distribution $P(\overline{w}, \bar{s})$ as:

$$P(\overline{w}, \bar{s}) = P_n(\bar{s})|\overline{w}|! \prod_{i=1}^{N} \frac{a_i^{w_i}}{w_i!} \tag{1}$$

where the unknown vector $P_n(\bar{s})$ represents the stationary probability distribution over the server states $\bar{s}$, given that the system contains n jobs (in service plus in queue). A key issue in the remainder of this section will be how to derive an expression for the $d(N, k)$-vector $\mathbf{P}_n$ with components $P_n(\bar{s})$. Here we use the shorthand notation $d(N, k) = (N + k - 1)!/(k!(n - 1)!)$. For example, for $N = 3$ and $k = 4$ we find $d(N, k) = 15$. Of course, this dimension increases rapidly with N and k. Now it is easy to check that in vector-notation:

$$\big((1 + \rho)\mathbf{I} + \bar{\mathbf{d}}\big)\mathbf{P}_n = \rho\mathbf{P}_{n-1} + \mathbf{A}\mathbf{P}_{n+1}. \tag{2}$$

Here $\mathbf{I}$ denotes the identity matrix and $\bar{\mathbf{d}}$ represents a diagonal matrix with elements $\delta(\bar{s})$ on the diagonal, i.e. $\bar{\mathbf{d}}\xi[\bar{s}] = \delta(\bar{s})\xi[\bar{s}]$, and the matrix $\mathbf{A}$ is defined by its working on an arbitrary $d(N, k)$-vector ξ, as:

$$\mathbf{A}\xi[\bar{s}] \stackrel{\text{def}}{=} \frac{1}{k} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i(1 + \delta_j)(s_j + 1 - e_{ij})\xi[\bar{s} - \bar{e}_i + \bar{e}_j].$$

The equation for $\mathbf{P}_n$ is a second order difference equation in a $d(N, k)$-dimensional linear space, which plays a central role in our analysis.

Note that $\mathbf{P}_n$ can also be defined for $n \leqslant k$. It is a vector of dimension $d(N, n)$, since only n of the servers are occupied. Let L_n denote the corresponding linear space of dimension $d(N, n)$ accommodating that vector. The equations for $n \leqslant k$ are

$$\mathbf{D}_n\mathbf{P}_n = \rho\mathbf{F}_n\mathbf{P}_{n-1} + \mathbf{B}_n\mathbf{P}_{n+1} \tag{3}$$

where $\mathbf{D}_n$, $\mathbf{F}_n$ and $\mathbf{B}_n$ are respectively defined as:

$$\begin{aligned}
&\mathbf{D}_n : \mathcal{L}_n \to \mathcal{L}_n, \quad \mathbf{D}_n\xi[\bar{s}] \stackrel{\text{def}}{=} \left(\frac{n}{k} + \rho + \delta(\bar{s})\right)\xi[\bar{s}], \\
&\mathbf{F}_n : \mathcal{L}_{n-1} \to \mathcal{L}_n, \ \mathbf{F}_n\xi[\bar{s}] \stackrel{\text{def}}{=} \sum_{i=1}^{N} a_i\xi[\bar{s} - \bar{e}_i], \\
&\mathbf{B}_n : \mathcal{L}_n \to \mathcal{L}_{n+1}, \ \mathbf{B}_n\xi[\bar{s}] \stackrel{\text{def}}{=} \frac{1}{k} \sum_{i=1}^{N} (s_i + 1)(1 + \delta_i)\xi[\bar{s} + \bar{e}_i], \quad \text{for } n < k.
\end{aligned}$$

For $n = k$, we get $\mathbf{B}_n = \mathbf{A}$ operating between $d(N, k)$-dimensional linear spaces. We come back to solving these equations in section 5.

4. The unperturbed case of an MCMS queueing system

Our analysis of multi-class multi-server problems starts by revisiting the well-known $M/M/k$ queue that can be considered as a special case where the service rate of each class is identical.

It is well known that the stationary probability distribution over n satisfies:

$$P(n) = \max\left(1, \frac{k}{n}\right)\rho P(n-1)$$

and the solution is found in a straightforward way. Now in case all service times are equal ($\delta_i = 0$), the multi-class probabilities $P(\overline{w}, \bar{s})$ are just a multinomial modification of the previously found $P(n)$, as we saw in the previous section. This can be checked by substitution in the steady state equations.

But, a lot more information about other solutions of the state equations for $n > k$ can be derived. This is useful for the exploration of the general case in the next section. First, we observe that the eigenvalues and eigenvectors of the matrix $\mathbf{A}$ can be explicitly found if all $\delta_i = 0$. We will come back to that shortly. Moreover, in this case ($\delta_i = 0$), the matrix $\bar{\mathbf{d}}$ in equation (2) vanishes and the solutions of this equation are then immediately found from the eigenvalues and eigenvectors of the matrix $\mathbf{A}$. That is, if we put $\mathbf{P}_n = z^{-(n-k)}V$ for some z with V equal to an eigenvector of $\mathbf{A}$ for the eigenvalue v, then we find a solution for $n > k$ if

$$\begin{aligned}(1+\rho)z^{-(n-k)}V &= \rho z^{(-n-1-k)}V + z^{-(n-k)}\mathbf{A}V \quad \Rightarrow \\ (1+\rho)zV &= \rho z^2 V + vV \quad \Rightarrow \quad (1+\rho)z = \rho z^2 + v.\end{aligned}$$

The eigenvectors and eigenvalues of the matrix $\mathbf{A}$ can be constructed as follows. First, we observer that

$$V_{(k,0,\ldots,0)}[\bar{s}] = k! \prod_{i=1}^{N} \frac{(a_i)^{s_i}}{s_i!}$$

defines an eigenvector of $\mathbf{A}$ for the eigenvalue 1. Let us interpret it as an eigenvector where all k servers are occupied with "real" jobs. In an analogous way, we can find eigenvectors $V_{(m,h_2,\ldots,h_N)}$ for eigenvalues m/k where only m servers are occupied with "real" jobs and $k-m$ are filled "artificially". Index h_j refers to the number of servers occupied with "virtual jobs" in "virtual mode j" with $j = 2, \ldots, N$.

Mathematically, this means that we decompose each state $\bar{s} = (s_1, \ldots, s_N)$ with $|\bar{s}| = k$ as

$$\begin{aligned}&s_i = \sum_{j=1}^{N} \theta_i^j, \ \theta_i^j \geqslant 0, \\ &\sum_{i=1}^{N} \theta_i^j = h_j, h_1 = m, \qquad \sum_{j=1}^{N} h_j = k.\end{aligned}$$

This allows for many possible decompositions θ. We define

$$V_{\bar{h}}[\bar{s}] = k! \sum_{\theta} \prod_{j=1}^{N} \prod_{i=1}^{N} \frac{(t_i^j)^{\theta_i^j}}{\theta_i^j!}. \tag{4}$$

Here we define the real mode $t^1 = (a_1, \ldots, a_N)$ and the virtual mode of type j as t^j with $t_j^j = -1$ and $t_i^j = 1/(N-1)$ for $i \neq j$. Note that for $j = 2, \ldots, N$ the vector t^j is perpendicular to $(1, \ldots, 1)$.

Again, one can check these eigenvectors by substitution using the definition of **A**. The eigenvalues corresponding to these eigenvectors are completely defined by the first component of the vector $\bar{h}$ which we denote as $m = h_1$. Hence, there are $k+1$ eigenvalues $v = m/k$ for $m = 0, \ldots, k$. The eigenspaces corresponding to each eigenvalue of **A** have dimension

$$d_{m/k} = \binom{N-2+k-m}{k-m}$$

due to the amount of possible combinations of the last $N-1$ components of the vectors $\bar{h}$, given that the first component is equal to m.

Then, by substituting the obtained eigenvalues $v = m/k$ for $m = 0, \ldots, k$ into the equation $(1+\rho)z_m = \rho z_m^2 + v$ we find the following values of z_m:

$$z_m = \frac{1}{2\rho}\left\{(1+\rho) \pm \sqrt{(1+\rho)^2 - 4\frac{m}{k}\rho}\right\}, \quad m = 0, \ldots, k.$$

Note that the $+$ sign leads to $z_m > 1$ and the $-$ sign to $z_m \leqslant 1$. Only the solutions with $z_m > 1$ decay for $n \to \infty$ and are acceptable in constructing probability distributions. There are $d(N,k)$ of such solutions. Only one of them, namely with $m = k$, plays a role in the exact solution for the unperturbed case. In this respect, the unperturbed case turns out to be special.

Of course, one should expect that this changes in a general case with $\delta_i \neq 0$.

5. Solving general MCMS problems

Let us first construct the solutions of the state equations for $n > k$ in general (section 5.1). Then we solve the remaining equations for $n \leqslant k$ and, thereby, construct the full solution of the state equations (section 5.2). Next, in section 5.3, we consider some special cases. Finally, we give some fast approximations of the exact solution in section 6.

5.1. The exact solution for $n > k$

Let us first reformulate the second order difference equation for the $d(N,k)$-vector $\mathbf{P}_n$ given in (2) as a first order difference equation for a $2d(N,k)$-dimensional vector. Now

let us consider the vector $\mathbb{P}_n = (\mathbf{P}_{n-1}, \mathbf{P}_n)^{\mathrm{T}}$. It has to satisfy $\mathbb{P}_n = \mathbb{H}\mathbb{P}_{n+1}$ with the matrix $\mathbb{H}$ given by

$$\mathbb{H} = \begin{pmatrix} \frac{1}{\rho}\big((1+\rho)\mathbf{I} + \mathbf{d}\big) & -\frac{1}{\rho}\mathbf{A} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}.$$

Note that solving this backward recursion boils down to determining the eigenvalues and eigenvectors of the matrix $\mathbb{H}$. The following information is crucial.

Lemma 1.

1. Under the stability condition with all δ_i sufficiently small, $\mathbb{H}$ has:
   - eigenvalues z satisfying $|z| \leqslant 1$ with total multiplicity $d(N,k)$,
   - eigenvalues z satisfying $|z| > 1$ with total multiplicity $d(N,k)$.
2. If $\delta_i \neq 0$, $\mathbf{A}$ has eigenvalues $\alpha(m/k)$, with $\alpha = \sum_{i=1}^{n} a_i(1+\delta_i)$, $m = 0, \ldots, k$, with the same multiplicities and similar eigenspaces as in the unperturbed case. As a consequence, 0 is an eigenvalue of $\mathbb{H}$ with eigenspace $(0, \ker(\mathbf{A}))^{\mathrm{T}}$ and the same multiplicity $d(N-1,k)$ as before.
3. $N+1$ special eigenvalues of $\mathbb{H}$ corresponding with eigenvectors possessing a product structure can be found:
   - there is an eigenvalue 1 with eigenvector $(\mathbf{B}, \mathbf{B})^{\mathrm{T}}$ with:

$$\mathbf{B}(\bar{s}) = |\bar{s}|! \prod_{i=1}^{N} \frac{a_i^{s_i}}{s_i!(1+\delta_i)^{s_i}}, \tag{5}$$

   - under the stability condition, there are N eigenvalues

$$z(\eta) = \frac{1+\rho-\eta}{\rho} > 1 \tag{6}$$

   where η satisfies a polynomial equation of degree $N+1$:

$$\frac{1+\rho-\eta}{\rho} = \sum_{i=1}^{N} a_i \frac{1+\delta_i}{\eta+\delta_i} \tag{7}$$

   which also has the solution $\eta = 1$, i.e. $z(1) = 1$ as above, see figure 1. The eigenvector for $z(\eta)$ is $(z(\eta)\mathbf{C}, \mathbf{C})^{\mathrm{T}}$ with:

$$\mathbf{C}(\bar{s}) = |\bar{s}|! \prod_{i=1}^{N} \frac{a_i^{s_i}}{s_i!(1+\delta_i/\eta)^{s_i}}. \tag{8}$$

One of these eigenvalues $z(\eta_0)$ crosses $z = 1$ into $z < 1$ crosses 1 into the region $\rho > 1$ where the stability condition is violated.

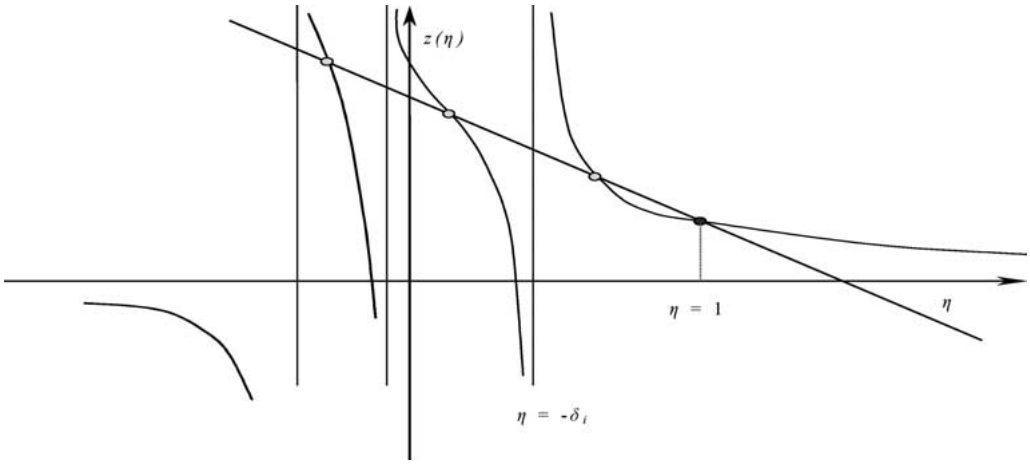

Figure 1. Graphical solution of the polynomial equation for some special eigenvalues of $\mathbb{H}$.

Proof. To start with we note that $z = 1$ is an eigenvalue of $\mathbb{H}$ corresponding with the given eigenvector. This follows simply by substitution. Note that in the eigenvector the probabilities over the states are proportional to the products of the required service fractions. Now the first part of the lemma is simply a consequence of the well-known perturbation theory for eigenvalues of matrices, cf. [8]. Next, we observe that also in the perturbed case the eigenvalues of $\mathbf{A}$ are $\alpha(m/k)$, with $m = 0, \ldots, k$, but now the eigenspaces are slightly different. They are found using a N-vector perpendicular to $\mathbf{1} + \boldsymbol{\delta} = (1 + \delta_1, \ldots, 1 + \delta_N)^{\mathrm{T}}$ instead of the expressions in section 4. Hence, the dimension of the eigenspaces corresponding with m/k is the same as before.

The remainder of the lemma is a matter of substituting the specified eigenvectors in the eigenvalue equation and checking that it is satisfied. The behaviour of the crucial eigenvalue follows by noticing that both the derivative of the left- and right-hand side at $\eta = 1$ are equal to $-\rho$. The situation is illustrated below. □

Let us now discuss the consequences of the lemma 1. First, it should be noted that in the case $k = 1$ (a) there are $N - 1$ eigenvalues 0 for any N, (b) there is an eigenvalue 1, and (c) there are N real eigenvalues >1 given by the special polynomial equation. This provides complete information on the eigenvalues. In other cases, other eigenvalues besides the special ones play a role. In figure 2 below, it is shown how the eigenvalues evolve from their unperturbed values if the perturbation is "turned on."

To plot the changes of eigenvalues, we use a case with $N = 4,\ k = 3,\ \rho = 0.9$. To change δ_i, we put $\delta_i = a_i t$ with $a_2 = 0.1,\ a_3 = 0.15,\ a_4 = 0.2$, when δ_1 is chosen such, that $\sum_{i=1}^{N} (a_i/(1 + \delta_i)) = 1$.

Let us now use the information on the eigenvalues and eigenspaces of $\mathbb{H}$ to solve the state equations for $n \geqslant k$ exactly in terms of the state probabilities for $n = k$. This can be done using the following recipe. Let us use the abbreviated notation $d = d(N, k)$.

Introduce:

- $\mathbf{\Omega}$ as the $d \times d$ diagonal (or Jordan) matrix corresponding with the eigenvalues of $\mathbb{H}$ with $|z| > 1$.
- $\mathbf{\Xi}$ as the $d \times d$ matrix of upper parts of the corresponding (generalised) eigenvectors.

Note that $\mathbf{Z} = \mathbf{\Xi}\mathbf{\Omega}\mathbf{\Xi}^{-1}$ satisfies

$$(1+\rho)\mathbf{I} + \bar{\mathbf{d}} = \mathbf{A}\mathbf{Z}^{-1} + \rho\mathbf{Z}.$$

Now

$$\mathbf{P}_n = \left(\mathbf{Z}^{-1}\right)^{n-k}\mathbf{P}_k = \mathbf{X}\mathbf{W}^{n-k}\mathbf{X}^{-1}\mathbf{P}_k$$

is an exact solution for $n \geqslant k$ starting at $\mathbf{P}_k$ for $n = k$. But, of course, $\mathbf{P}_k$ for $n = k$ still has to be determined by analysing the equations for $n \leqslant k$.

5.2. *The exact solution for $n \leqslant k$*

Consider $\mathbf{P}_n(\bar{s})$ for $n < k$ satisfying (3). Again, the state equations for $n < k$ have a backward recursion structure. The equation for $n = k$

$$\left((1+\rho)\mathbf{I} + \bar{\mathbf{d}}\right)\mathbf{P}_n = \rho\mathbf{P}_{n-1} + \mathbf{A}\mathbf{P}_{n+1}$$

reduces to

$$\mathbf{P}_k = \mathbf{Z}^{-1}\mathbf{F}_k\mathbf{P}_{k-1}$$

by using the equation for $\mathbf{Z}$. For $n < k$, we obtain

$$\mathbf{P}_n = \rho\mathbf{D}_n^{-1}\mathbf{F}_n\mathbf{P}_{n-1} + \mathbf{D}_n^{-1}\mathbf{B}_n\mathbf{P}_{n+1}.$$

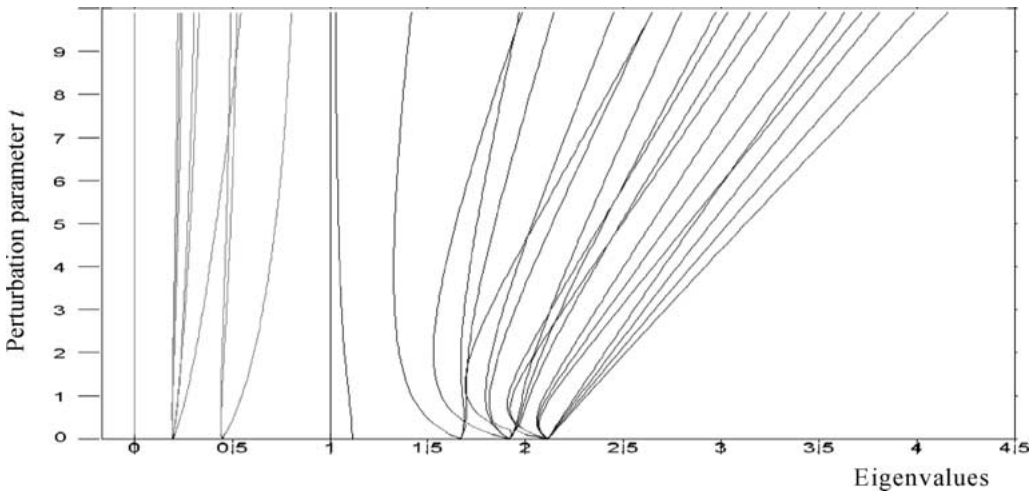

Figure 2. A sketch of the behaviour of the eigenvalues of $\mathbb{H}$ with the strength of the perturbation.

Hence, the complete solution can be represented as

$$\mathbf{P}_n = \mathbf{Q}_n \mathbf{P}_{n-1} = \mathbf{Q}_n \mathbf{Q}_{n-1} \cdots \mathbf{Q}_1 \mathbf{P}_0$$

where $\mathbf{Q}_n$ follows recursively from

$$\mathbf{Q}_k = \mathbf{Z}^{-1} \mathbf{F}_k$$

and

$$\mathbf{Q}_n = \rho(\mathbf{D}_n - \mathbf{B}_n \mathbf{Q}_{n+1})^{-1} \mathbf{F}_n.$$

Of course, $\mathbf{P}_0$ is a free constant determined by

$$\sum p(\overline{w}, \bar{s}) = 1.$$

A simple computation shows that

$$\mathbf{P}_0 = \left\{1 + \langle \mathbf{1}_1, \mathbf{Q}_1 \rangle_1 + \cdots + \langle \mathbf{1}_{k-1}, \mathbf{Q}_{k-1} \cdots \mathbf{Q}_1 \rangle_{k-1} + \left\langle \mathbf{1}_k, \left(\mathbf{I} - \mathbf{Z}^{-1}\right)^{-1} \mathbf{Q}_k \cdots \mathbf{Q}_1 \right\rangle_k \right\}^{-1}.$$

Here we denote with $\langle \mathbf{1}_n, \mathbf{X} \rangle_n$ the inner product of $\mathbf{X}$ in a $d(N, n)$-dimensional space with the $d(N, n)$-vector $\mathbf{1}_n$ with all components equal to 1.

Thus, the exact solution for all n has now been derived.

5.3. Some special cases

To illustrate the theory, we consider two cases where the algorithm can be executed more explicitly: (I) the case $k = 1$, (II) the case $k = 2,\ N = 2$.

In the case $k = 1$, we have states $s = e_1, \ldots, e_N$ indicating the type of job in execution. The eigenvalues are given by $z(\eta_j),\ j = 1, \ldots, N$, as defined by equations (6), (7); the matrix of eigenvectors $\mathbf{\Xi}$ has elements $a_i/(1 + \delta_i/\eta_j)$. The solution of (2) is given by: $\mathbf{P}_n = \mathbf{P}_0 \mathbf{Z}^{-n} \mathbf{a}$. Here $\mathbf{a}$ denotes the N-vector with components a_i. In the case of 2 classes this coincides with a result that can be found in [9].

In case $k = 2$ and $N = 2$, we start with the observation that the special eigenvalues in equations (6), (7) can be explicitly computed, because the degree of the equation is 3 and the solution $\eta = 1$ is already known. This leads us to the following quadratic equation for z

$$0 = eq_1(z) \stackrel{\text{def}}{=} \rho^2 z^2 - \rho\{2 + \rho + \delta_1 + \delta_2\} z + \left\{1 + \rho + \delta_1 + \delta_2 + (1 - \rho)\delta_1 \delta_2\right\}.$$

Now, only 2 of the 6 eigenvalues of $\mathbb{H}$ are still unknown; they can be found from the characteristic polynomial $0 = \mathbf{eq}(z) \stackrel{\text{def}}{=} \det(-(1 + \rho + \mathbf{D})z + \rho z^2 \mathbf{I} + \mathbf{A})$ by dividing out the known eigenvalues. Here we denote by $\mathbf{D}$ the diagonal matrix with the average perturbation δ for the respective states on the diagonal. Hence, we decompose $\mathbf{eq}(z) = z(z - 1) eq_1(z) eq_2(z)$. Therefore, the remaining 2 eigenvalues have to satisfy a quadratic

equation with coefficients that can be explicitly given in terms of the original matrix coefficients as

$$0 = eq_2(z) \stackrel{\text{def}}{=} \rho z^2 - \rho\left\{1 + \rho + \frac{1}{2}(\delta_1 + \delta_2)\right\}z + \frac{1}{2}\{1 + a_1\delta_1 + a_2\delta_2\}.$$

Note that the latter equation gives rise to one eigenvalue $z > 1$.

Now $\mathbf{Z}$ is a 3×3 matrix operating on the probabilities of the states (2, 0), (1, 1), (0, 2) for the jobs in execution when the servers are occupied. The column of Ξ corresponding with an eigenvalue $z = (1 + \rho - \eta)/\rho$ becomes $(v, 1, 1/v)^{\mathrm{T}}$ with $v = a_1(\eta + \delta_2)/\{a_2(\eta + \delta_1)\}$. For the other eigenvector corresponding with $z > 1$ satisfying $eq_2(z) = 0$, we find $-a_1(1 + \delta_2)/q(z), 1, a_2(1 + \delta_1)/q(z)^{\mathrm{T}}$ with $q(z) = \{(\delta_2 - \delta_1)/2\}z + \{a_1(1 + \delta_1) - a_2(1 + \delta_2)\}/2$. It is clear that also in this case the solution can be given explicitly in terms of the parameters of the problem.

6. Approximate solutions

In this section we shall introduce some approximations of the state probabilities $P_n(\bar{s})$. These approximations will be constructed using only the eigenmodes of the steady state equation for $n > k$, which have a multiplicative structure. However, the equation for $n \leqslant k$ the equations can only be satisfied if $k = 1$. In the general case they will be satisfied only in an approximate way.

The advantage of the approximations is that (I) the computation of the approximation requires algebra in spaces of considerably lower dimension and (II) all sums over states in the performance measures can be determined explicitly due to the product structure. Moreover, since the critical mode which causes instability, if the utilisation crosses 1, is included, these approximations can be considered as heavy traffic approximations. For a further discussion on heavy traffic approximations we refer to [2,18]. The approximation is given by:

$$\text{for } n < k, \quad P_n(\bar{s}) = C(k\rho)^n \prod_{i=1}^{N} \frac{a_i^{s_i}}{s_i!(1 + \delta_i)^{s_i}}, \tag{9}$$

$$\text{for } n \geqslant k, \quad P_n(\bar{s}) = \sum_{i=1}^{N} \gamma_i z_i^{-(n-k)} D(\eta_i)^{-k} |\bar{s}|! \prod_{j=1}^{N} \frac{a_j^{s_j}}{s_j!(1 + \delta_j/\eta_i)^{s_j}} \tag{10}$$

with η_i, z_i as defined in (6), (7) and

$$D(\eta_i) = \sum_{j=1}^{N} \frac{a_j}{(1 + \delta_j/\eta_i)}.$$

In order to determine the constants C and γ_i we require that the total probability sums to 1 and that the expected number of items of type i in service is correct for each i. This provides us with $N+1$ equations for $N+1$ unknowns:

$$\sum_{i=1}^{N}\left\{D(\eta_i)^{-1}\frac{a_m}{(1+\delta_m/\eta_i)}\right\}\varphi_i + Cb_m\sum_{n=1}^{k-1}\frac{n}{k}\frac{(k\rho)^n}{n!} = \rho_m,$$
$$\sum_{i=1}^{N}\varphi_i + C\sum_{n=0}^{k-1}\frac{(k\rho)^n}{n!} = 1$$

with $\varphi_i \stackrel{\text{def}}{=} \gamma_i(1-z_i^{-1})^{-1}$ and hence

$$C = \left\{\sum_{n=0}^{k-1}\frac{(k\rho)^n}{n!}\left(1-\frac{n}{k}\right)\right\}^{-1}(1-\rho).$$

An even simpler approximation is found, if we use for $n > k-1$ only the critical mode corresponding with the least decay for increasing n. In this case, we require the expected number in service disregarding type to be correct and we obtain the same value for C as given here above, but now

$$\gamma_{cr} = \left(1-z_{cr}^{-1}\right)\left\{1 - C\sum_{n=1}^{k-1}\frac{(k\rho)^n}{n!}\right\}$$

and the other coefficients vanish. In the next section we shall compare these approximations with some well-known heuristics given by Whitt [17].

7. Accurate information on performance measures

Here we shall provide some further insight in the behaviour of some interesting performance measures related to stochastic variables, such as the waiting time w, the queue length q and the number of items of type i in queue q_i. Performance measures that we consider are expectations, variances and probability for positivity of these stochastic variables. Moreover, since it is clear that there is correlation between these stochastic variables, we also shall provide formulas for some conditional expectations and correlation coefficients. Of course, the clue that such exact formulas can easily be derived lies in the fact that series of geometric type can simply be summed. To start, we notice that

$$P(q>0) = \left\langle \mathbf{1}_k, (\mathbf{Z}-\mathbf{I})^{-1}\mathbf{P}_k\right\rangle,$$
$$P(w>0) = \left\langle \mathbf{1}_k, (\mathbf{Z}-\mathbf{I})^{-1}\mathbf{Z}\mathbf{P}_k\right\rangle,$$
$$P(q_i>0) = \left\langle \mathbf{I}_k^i, (\mathbf{Z}-\mathbf{I})^{-1}\mathbf{P}_k\right\rangle$$

and the performance estimators for the total number of items in queue are

$$E[q] = \left\langle \mathbf{1}_k, (\mathbf{Z}-\mathbf{I})^{-2}\mathbf{Z}\mathbf{P}_k\right\rangle,$$

$$E\big[q^2\big] = E\big[q(q-1)\big] + E[q] = 2\big\langle \mathbf{1}_k, (\mathbf{Z}-\mathbf{I})^{-3}\mathbf{Z}\mathbf{P}_k\big\rangle,$$
$$\mathrm{Var}[q] = E\big[q^2\big] - E[q]^2.$$

Then we can write down the performance estimators for the number of items in queue for each type of items:

$$E[q_i \mid q = n-k] = a_i(n-k),$$
$$E[q_i] = a_E[q],$$
$$E\big[q_i(q_i-1) \mid q = n-k\big] = a_i^2(n-k)(n-k-1),$$
$$E\big[q_i(q_i-1)\big] = a_i^2 E\big[q(q-1)\big],$$
$$\mathrm{Var}[q_i] = E\big[q_i(q_i-1)\big] + E[q_i] - E[q_i]^2.$$

We can also estimate the covariance between the numbers of items of different classes in queue:

$$E[q_i q_j \mid q = n-k] = a_i a_j (n-k)(n-k-1) \text{ for } i \neq j,$$
$$E[q_i q_j] = a_i a_j E\big[q(q-1)\big] \qquad \text{for } i \neq j,$$
$$corr(q_i, q_j) = E[q_i q_j] - E[q_i]E[q_j].$$

Here we use the notation $\mathbf{1}_n$ as before in determining $\mathbf{P}_0$. The matrix $\mathbf{I}_n^i$ is a diagonal matrix of dimension $d(N,n)$ with 1 for each state $\bar{s}$ that contains type i and 0 elsewhere. Note that Little's law is satisfied and the expected waiting time does not depend on the type of item.

Some other quantities are simpler to derive. Using Little's theorem, we find the first moments:

$$E[R_i] = E[q_i] + \frac{\lambda_i}{\mu_i},$$
$$E[S_i] = \frac{\lambda_i}{\mu_i}.$$

Moreover, we can now easily derive the following second moments and correlation coefficients:

$$E\big[R_i^2\big] = E\big[q_i^2\big] + E\big[S_i^2\big] + 2E[S_i q_i],$$
$$E\big[S_i^2\big] = \sum_{n=1}^{k-1} \big\langle \chi_n^{s_i^2}, \mathbf{P}_n\big\rangle + \big\langle \chi_k^{s_i^2}, \big(\mathbf{I}-\mathbf{Z}^{-1}\big)^{-1}\mathbf{P}_k\big\rangle,$$
$$E[S_i q_j] = a_j\big\langle \chi_k^{s_i}, \big(\mathbf{I}-\mathbf{Z}^{-1}\big)^{-2}\mathbf{Z}^{-1}\mathbf{P}_k\big\rangle,$$
$$E[S_i S_j] = \sum_{n=1}^{k-1} \big\langle \chi_n^{s_i \cdot s_j}, \mathbf{P}_n\big\rangle + \big\langle \chi_k^{s_i \cdot s_j}, \big(\mathbf{I}-\mathbf{Z}^{-1}\big)^{-1}\mathbf{P}_k\big\rangle,$$
$$E[R_i R_j] = E[S_i S_j] + E[S_i q_j] + E[S_j q_i] + E[q_i q_j].$$

In these formulas, $\chi_n^{s_i^2}$ and $\chi_n^{s_i \cdot s_j}$ are the diagonal matrices of dimension $d(N, n)$ with the square number of items of type i in $\bar{s}$ (i.e. s_i^2) or, accordingly, the product of the numbers of types i and j items in $\bar{s}$ (i.e. $s_i \cdot s_j$) as the diagonal element corresponding to $\bar{s}$.

In an analogous way, we can analyse different performance measures in stochastic spare parts systems (cf. [15]). For example, mean and variances of the number of backorders for type i, $BO_i = \max(0, R_i - st_i)$ with st_i a given positive integer representing the stock level in spare part networks. It is possible using recursion with respect to st_i, since first and second moments and the correlation coefficients of the backorder variables with stock levels $st_i = 0$ are known. This basic idea can already be found in [14]. In our work we extend it to include correlation between backorders of different types. This provides us with the possibility to compute a better approximation of the system availability in Sherbrooke's spare parts stock allocation method METRIC. For details we refer again to [15]. Some results will be shown in the sequel.

8. Some numerical results

In order to derive numerical results, we implemented the exact solutions and the approximations as described above in MATLAB. A wealth of interesting phenomena can now be quantitatively analysed. Let us just show a few of the results.

8.1. The average waiting time in MCMS systems and correlation between different types of items

The effects of different service times for various classes of items can easily be illustrated. Consider a system with 2 classes and k servers. The parameters are chosen as $a_1 = 1/3$, $a_2 = 2/3$. The first type of items has an average service time larger than the second type of items. In terms of δ_1 and δ_2 this can be represented as:

$$\delta_1 = -\delta, \qquad \delta_2 = \frac{(1/2)\delta}{1 - (2/3)\delta}, \quad \text{with } \delta \in \left[0, \frac{2}{3}\right), \quad \text{i.e.} \quad \sum_{i=1}^{2} \frac{a_i}{1 + \delta_i} = 1.$$

Let us now compare for various choices of ρ and k the expected waiting time in the system with $\delta > 0$ with those for $M/M/k$ where $\delta = 0$ (figure 3). The conclusion that the ratio increases with δ is not surprising, but it is nice that we can explicitly determine how it increases. Of course, a similar behaviour as in figure 3 is found for the average number of items in the system.

8.2. Variance per item and correlation between different types of items

Of course, a similar behaviour as in figure 3 is found for the average number of items in the system since the number of items in service is constant. It is also interesting to notice that the variance of the number of items in the system is also increases with increase of δ. In figure 4 we have presented the relative increase of the variance of the total number of items in the system.

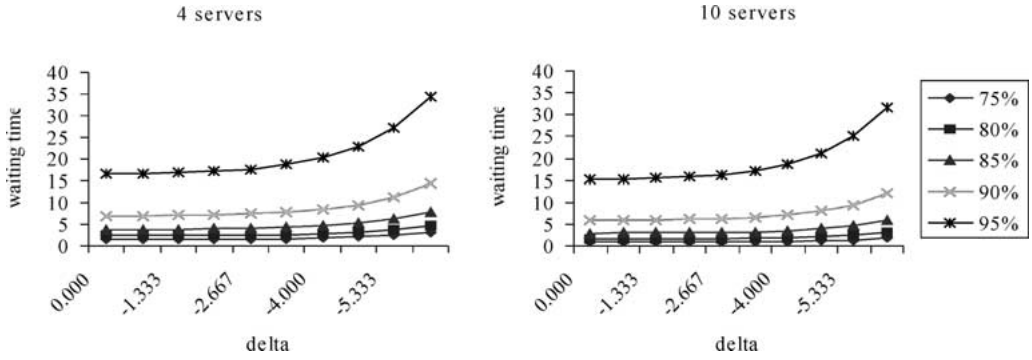

Figure 3. Increasing differences between service characteristics of classes lead to increasing waiting times.

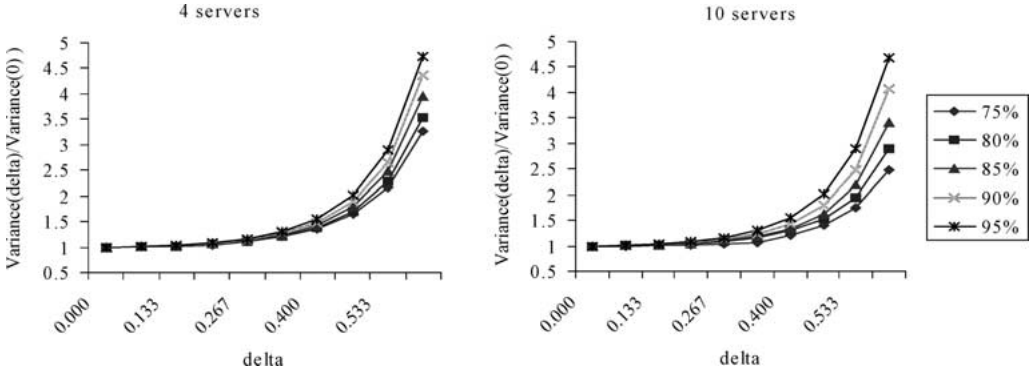

Figure 4. Increasing differences between service characteristics of classes lead to increasing variance of items in the system.

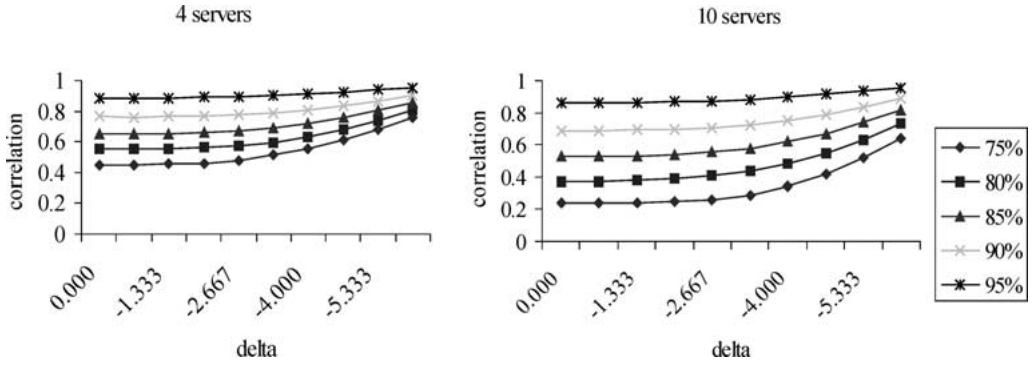

Figure 5. Increasing differences between service characteristics of classes lead to increase of the correlation coefficient.

Let us again consider the same example, but now with our focus on the interdependency between stochastic variables for different types of items. In figure 5, the behaviour of the correlation coefficient is shown.

From figure 5, we see that the correlation between backorder levels can be significant. Even at a moderate utilisation (75%), the correlation coefficient may exceed 0.5. For high utilisation ($\rho = 0.95$), it is clear that the correlation is very high (around 0.9), suggesting that it is relevant to take the correlations into account when estimating the system availability. In the next subsection, we deal with this issue.

8.3. An application: better approximations for system availability

Applications of our results can be found in several areas. Here we consider the situation of an installed base of size B where upon failure items are repaired in a non-dedicated repair shop, while for item type i there are st_i spare parts. Let us first have a closer look at the asymptotic expansion of the nonlinear system availability function of VARI-METRIC model (cf. [14]). In their simplest form VARI-METRIC models use normally a linear approximation

$$A_1 = 1 - \frac{1}{B} \sum_i E\big[BO_i(\overline{st})\big].$$

Although this approximation seems adequate for infinite repair shop capacities, it is questionable whether this is also true in the case of *finite* capacities. The latter means that the number of backorders of different components at the same repair shop are mutually *correlated.* This is a severe complication, as the expectation of the product of backorders cannot be taken term-wise anymore. Still, equation (11) can be seen as a linear approximation of the exact availability function

$$A(\overline{st}) = E\left[\frac{1}{B} \prod_i \big\{1 - BO_i(\overline{st})\big\}\right].$$

The quality of this linear approximation is still unknown, however. The analysis of MCMS models provides insight in the *joint* probability distribution of the number in repair for all items. This gives us the expectations for backorders per item as well as the correlations between items. Therefore, it is possible to have an approximation where also the quadratic terms are taken into account:

$$A_2 = 1 - \sum_m E\big[BO_{mi}(\overline{st})\big] + \sum_{i,j,i<j} E\big[BO_i(\overline{st})BO_j(\overline{st})\big].$$

Note that the correction to the linear approximation is positive. Here we use exact expressions for the moments of the backorders. In figure 6, we show the difference between the two approximations. Let us take an example of a single-location, single indenture system with $k = 4$ servers and $N = 5$ types of items processed in the same repair shop. The arrival (failure) rates fractions are $a_i = 0.2$ for all i. For the service rate deviation δ_i, we put $0.5(i-1)$ for $i = 2, \ldots, 5$ and δ_1 is negative $= -21/31$ so that $\sum a_i/(1+\delta_i) = 1$. All spare part levels are equal to $st_i = 7$ items in stock. The size of the installed base B is 15. The utilisation rate is varied between $\rho = 0.6$ and $\rho = 0.84$.

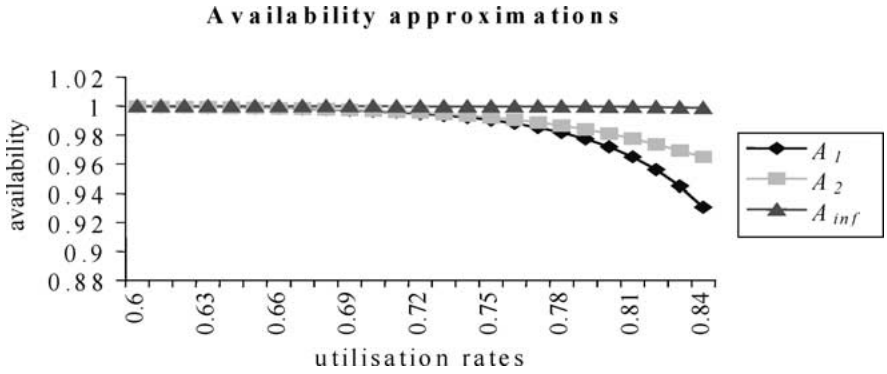

Figure 6. Approximations of the system availability.

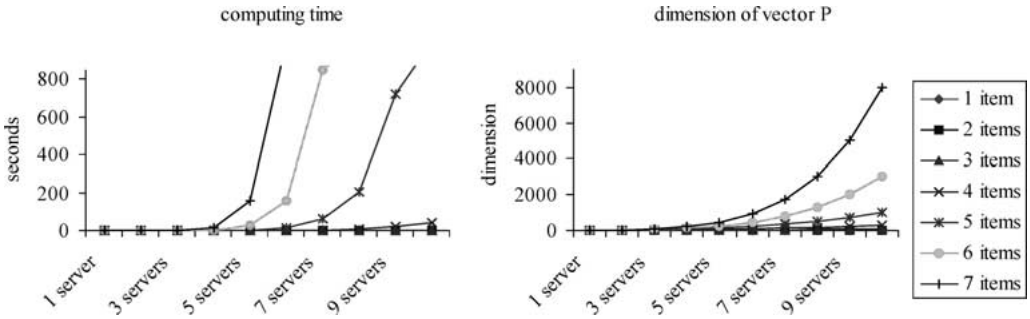

Figure 7. Computing times and dimensions of the vector **P** for different number of item classes in the system and different amounts of servers.

In figure 6, we have also shown another approximation for the system availability, denoted by A_{inf}. This estimate is obtained under the assumption of infinite repair capacity and based on a general service time distribution (waiting and repair together!) for items of type i, having a mean value R_i as derived in section 7. Using Palm's theorem, cf. [11], then the number of items of a type i in the repair shop has a Poison distribution with R_i as parameter. Different items have independent distributions. This infinite capacity approximation is often used in practice, cf. (Rustenburg, 2000). Note from the figure above that considerable differences between the approximations of the system availability arise for high utilisation rates.

8.4. Computational efforts

To estimate the computational efforts needed to solve the problem exactly, we estimated the CPU time usage for the most demanding computations, namely the calculation of the eigenvectors of the matrix $\mathbb{H}$. Below, we show some results for various values for the number of item classes N and the number of servers k. The computing time was estimated in seconds using MATLAB 5.3 with NAG toolbox, using a Pentium II-350 PC with 128 Mb RAM and under Windows NT.

Table 1
Computing time (sec.) required to find eigenvalues for different number of item classes in the system and different amounts of servers.

	1 ser.	2 ser.	3 ser.	4 ser.	5 ser.	6 ser.	7 ser.	8 ser.	9 ser.	10 ser.
2 classes	0	0	0	0	0	0	0	0	0	0.01
3 classes	0	0	0	0.01	0.02	0.05	0.09	0.16	0.271	0.41
4 classes	0	0	0.01	0.07	0.26	0.771	2.583	8.382	19.618	43.853
5 classes	0	0.01	0.07	0.43	2.484	15.752	60.387	201.38	720.55	–
6 classes	0	0.07	0.24	2.474	26.448	155.30	–	–	–	–

Table 2
Dimensions of the vector **P** for different number of item classes in the system and different amounts of servers.

	1 ser.	2 ser.	3 ser.	4 ser.	5 ser.	6 ser.	7 ser.	8 ser.	9 ser.	10 ser.
2 classes	2	3	4	5	6	7	8	9	10	11
3 classes	3	6	10	15	21	28	36	45	55	66
4 classes	4	10	20	35	56	84	120	165	220	286
5 classes	5	15	35	70	126	210	330	495	715	1001
6 classes	6	21	56	126	252	462	792	1287	2002	3003

The matrix $\mathbb{H}$ is first balanced and then reduced to upper Hessenberg form using real stabilised elementary similarity transformations. The eigenvalues and eigenvectors of the Hessenberg matrix are calculated using the QR algorithm. Next, the eigenvectors of the Hessenberg matrix are transformed back to the eigenvectors of the original matrix $\mathbb{H}$ (cf. [8,13]). The total computation time of both these algorithms is polynomial in dimension of the matrix $\mathbb{H}$ (figure 7, tables 1 and 2).

The highest dimension of the cases solved here ($d = 715$, for 5 items and 9 servers) seems already quite high for a practical use, and shows that this exact algorithm can be easily used in practice, certainly if a faster computer is used. Otherwise, the approximation scheme (section 6) can be applied.

8.5. On the quality of the approximate probability distribution

Finally, we shall provide some insight in the errors by using the simpler N mode or 1-critical mode approximations introduced in section 6. Let us take an example with $N = 5$ classes of items with arrival fractions $a_i = 0.2$ for all i. For the service rate deviation δ_i, we put $0.5(i-1)$ for $i = 2, \ldots, 5$, $\delta_1 = -21/31$, so that $\sum a_i/(1+\delta_i) = 1$, where $a_i/(1+\delta_i)$ represent the service fractions. The utilisation rate is varied between $\rho = 0.7$ and $\rho = 0.95$. The number of servers is varied between $k = 1$ and $k = 4$. In figure 8, the relative errors for the first and second moments are compared to the exact solution.

In these examples, with moderate ratios of service fractions for different classes and high utilisation rates, the errors are around 5%, which is quite reasonable. Of course, one should be careful with these approximations if the number of servers is large.

The approximations are better for high utilisation rates, since the smallest (critical) eigenvalue (>1) becomes closer to one and the influence of the others eigenvalues becomes negligible. This is known as state space collapse in heavy traffic approximations. For further discussions of this phenomenon, we refer to [2,18].

Comparison of the obtained approximations to the approximations for multi-server [17] modified by Diaz and Fu [7] for the multi-class case shows that the Whitt's and Diaz's approximations give the same results for the first moment as N-mode MCMS approximation. This can be understood from the fact that in both cases the expected waiting time for k servers is derived with a scaling technique from the exact result for the expected waiting time for 1 server. All compared approximations have negligible computation time (less than 10^{-3} s) for the presented experiments. However, the approximation of Diaz and Fu for the second moment for a single server queue (cf. [7]) produces errors up to 5%, while our results are exact. A comparison of the variance of the total number of items in the system obtained by our approximation and by Whitt's approximation for $GI/G/k$ queue is presented in (figure 9). There we can see that, although the Whitt's approximation is more stable to the number of classes, our approximation produces smaller errors. Therefore, our approximation, which can also give us other performance estimators (e.g., correlations, higher moments), is preferable for situations where this is important (such as in section 8.3 with N high).

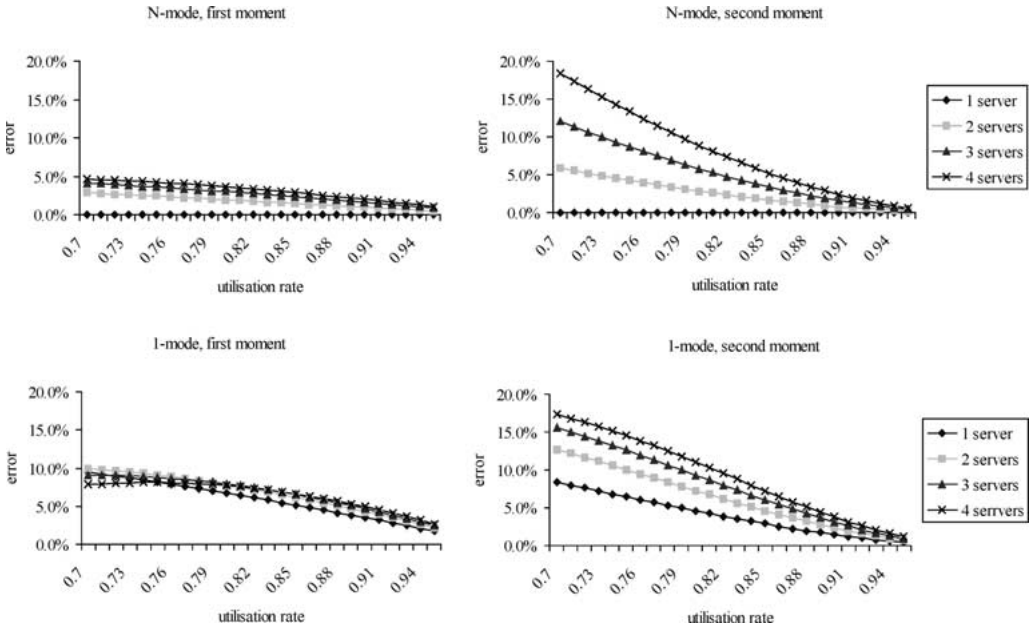

Figure 8. The approximate distributions are usually rather accurate for high utilisation rates.

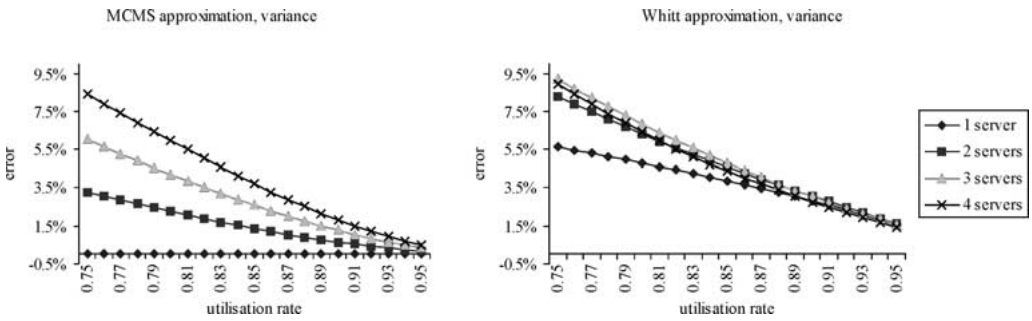

Figure 9. Average errors of the second[1] moment produced by N-mode MCMS approximation and Whitt's approximation.

9. Conclusions and generalisations

In this section, we derived a method for the exact analysis of multi-class, multi-sever queues, based on a classical method using the stationary state equations. Though conceptually we only deal with a perturbation of the well-known single class $M/M/k$ system, the structure of the solution becomes a lot more complex. Using the exact solution, several performance measures of the MCMS system can be studied in terms of formulas with a finite number of terms. The computational effort to find the exact solution depends on the number of classes N and the number of servers k. Representative for the computational effort is the dimension of the linear space used in section 5, i.e. $d(N,k) = (N+k-1)!/\{k!(N-1)!\}$. For large instances of the problem, some well-founded approximations can be given which only rely on N modes.

The exact method introduced in this section can in principle be generalised to problems with non-identical servers, but the computational effort will increase since we will have to keep track of state of each server and then we get $d(N,k) = N^k$. Also, certain situations with priority classes can be tackled in the spirit of this paper. This is still work in progress, cf. [15].

References

[1] D. Bertsimas and G. Mourtzinou, Multiclass queueing systems in heavy traffic: An asymptotic approach based on distributional and conservation laws, Oper. Res. 45(3) (1997) 470–487.
[2] M. Bramson, State space collapse with application to heavy traffic limits for multiclass queueing networks, Queueing Systems 30 (1998) 89–148.
[3] J.A. Buzacott and J.G. Shanthikumar, *Stochastic Models of Manufacturing Systems* (Prentice-Hall, Englewood Cliffs, NJ, 1993).
[4] J.W. Cohen, On the $M/G/2$ queueing model, Stochastic Process. Appl. 12 (1982) 231–248.
[5] J.H.A. de Smit, The queue $GI/M/s$ with customers of different types or the queue $GI/H_{\mathrm{m}}/s$, Adv. in Appl. Probab. 15 (1983) 392–419.

[1] For the first moment of numbers of items in the system both approximations (N-mode and Diaz's) produce identical errors.

[6] J.H.A. de Smit, A numerical solution for the multi-server queue with hyper-exponential service time, Oper. Res. Lett. 2(5) (1983) 217–224.
[7] A. Diaz and M.C. Fu, Models for multi-echelon repairable item inventory systems with limited repair capacity, European J. Oper. Res. 97 (1997) 480–492.
[8] G.H. Golub and C.F. van Loan, *Matrix Computations*, 3rd ed. (John Hopkins Univ. Press, Baltimore and London, 1996).
[9] D. Gross and C.M. Harris, *Fundamentals of Queueing Theory* (Wiley, New York, 1998).
[10] R.W. Hall, *Queueing Methods for Services and Manufacturing* (Prentice-Hall, Englewood Cliffs, NJ, 1991).
[11] C. Palm, Analysis of the Erlang traffic formulae for busy-signal arrangements, Ericsson Technics 4 (1938) 39–58.
[12] H.T. Papadopolous, C. Heavey and J. Browne, *Queueing Theory in Manufacturing Systems Analysis and Design* (Chapman & Hall, London, 1993).
[13] Y. Saad, *Numerical Methods for Large Eigenvalue problems: Theory and Algorithms* (Willey, New York, 1992).
[14] C.C. Sherbrooke, *Optimal Inventory Modelling of Systems: Multi-Echelon Techniques* (Wiley, New York, 1992).
[15] A. Sleptchenko, M.C. van der Heijden and A. van Harten, Effects of finite repair capacity in multi-echelon, multi-indenture service part supply systems, Internat. J. Production Economics 79(3) (2002) 209–230.
[16] A. Sleptchenko, A. van Harten and M.C. van der Heijden, Analyzing multi-class, multi-server queueing systems with preemptive priorities, Working paper, University of Twente, Faculty of Technology and Management, Enschede, The Netherlands (2002) submitted for publication.
[17] W. Whitt, Approximations for the $GI/G/c$ queue, Production Oper. Managm. 2 (1993) 144–161.
[18] R.J. Williams, Diffusion approximations for open class queueing networks: Sufficient conditions involving state space collapse, Queueing Systems 30 (1998) 27–88.