EURANDOM PREPRINT SERIES
2010-004

A reversible loss system with multi-type customers and multi-type servers

I. Adan, C. Hurkens, G. Weiss
ISSN 1389-2355

A reversible loss system with multi-type customers and multi-type servers

Ivo Adan* Cor Hurkens* Gideon Weiss†

January 25, 2010

Abstract

We consider a memoryless loss system with servers $\mathcal{S} = \{1, \ldots, J\}$, and with customer types $\mathcal{C} = \{1, \ldots, I\}$. Servers are multi-type, so that server j can serve a subset of customer types $C(j)$. We show that the probabilities of assigning arriving customers to idle servers can be chosen in such a way that the Markov process describing the system is reversible, with a simple product form stationary distribution.

Keywords: Service system; loss system; multi type customers; multi type servers; product form solution; reversible Markov chain.

1 Model

We consider a loss system with servers $\mathcal{S} = \{1, \ldots, J\}$, and with customer types $\mathcal{C} = \{1, \ldots, I\}$. Arrivals are Poisson. Customers of type i arrive at rate λ_i. The service requirements of all customers are i.i.d. exponentially distributed with rate 1. Servers are multi-type, so that server j can serve a subset of customer types $C(j)$. Server j works at rate μ_j.

The system is a loss system: Customers that arrive, and do not find an idle server which can serve them, are lost. We define the state of the system at time t as $X(t) = S$, where $S \subseteq \mathcal{S}$ is the set of idle servers which are available to receive customers at time t.

To complete the description of the system we need to specify how arriving customers are assigned to servers: An arriving customer of type i which arrives when the system is in state S will choose server $j \in S$ (where $i \in C(j)$) with probability $P(i, j|S)$. With this assignment $X(t)$ is a continuous time finite state Markov chain (CTMC).

Loss systems with multi-type servers and multi-type customers are motivated by applications such as, e.g., call centers with skill based routing [6, 10], redundant data storage for video on demand [5] or bed capacity planning of hospital wards [7, 12].

Our result in this paper is to show that one can choose $P(i, j|S)$ in such a way that the Markov process $X(t)$ is reversible, and as a result, one can then write down the stationary distribution of the process explicitly. This stationary distribution is unique, even though $P(i, j|S)$ which lead to it may not be unique.

*Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, the Netherlands; email {iadan,wscor}@win.tue.nl

†Department of Statistics, The University of Haifa, Mount Carmel 31905, Israel; email gweiss@stat.haifa.ac.il
Research supported in part by Israel Science Foundation Grants 454/05 and 711/09

Example 1:

Let $\mathcal{S} = \{1, 2\}$, $\mathcal{C} = \{1, 2\}$ and $C(1) = \{1, 2\}$, $C(2) = \{1\}$. If a type 1 customer arrives in an empty system, then this customer is sent to server 1 or to server 2 with corresponding probabilities $P(1, 1|\{1, 2\})$, $P(1, 2|\{1, 2\})$. If we choose

$$P(1, 1|\{1, 2\}) = \frac{\lambda_1}{2\lambda_1 + \lambda_2}, \qquad P(1, 2|\{1, 2\}) = \frac{\lambda_1 + \lambda_2}{2\lambda_1 + \lambda_2}, \tag{1}$$

then the stationary distribution is:

$$\pi(\{1\}) = \pi(\emptyset) \frac{\mu_1}{\lambda_1 + \lambda_2}, \quad \pi(\{2\}) = \pi(\emptyset) \frac{\mu_2}{\lambda_1}, \quad \pi(\{1, 2\}) = \pi(\emptyset) \frac{\mu_1 \mu_2 (2\lambda_1 + \lambda_2)}{\lambda_1 (\lambda_1 + \lambda_2)^2}, \tag{2}$$

where $\pi(\emptyset)$ normalizes the sum to 1.

Notation:

To facilitate reading we will use index i for customer types, and indexes j, k for servers, and we will use S for subsets of servers, C for subsets of customer types. We will denote by $C(S) = \bigcup_{j \in S} C(j)$ the set of customer types which can be served by at least one server in S. We will also denote by $S(i)$ the set of servers that can serve customers of type i.

2 Reversibility and product form

We now show that the assumption of reversibility uniquely determines the transition rates of the CTMC, and induces a simple product form stationary distribution. The process $X(t)$ is reversible if and only if the CTMC $X(t)$ satisfies the detailed balance equations (see Theorem 1.2 in [11]).

We denote by $\eta_j(S)$ the rate at which server $j \in S$ becomes busy, when the system is in state S. Detailed balance equations for the stationary probabilities $\pi(S)$ hold if:

$$\pi(S)\, \eta_j(S) = \pi(S \backslash \{j\})\, \mu_j, \quad \text{for all subsets } S \text{ and } j \in S \tag{3}$$

If detailed balance (3) holds, we get for $S = \{j_1, \ldots, j_m\}$:

$$\pi(S) = \pi(\emptyset)\, \frac{\mu_{j_1}}{\eta_{j_1}(\{j_1\})} \frac{\mu_{j_2}}{\eta_{j_2}(\{j_1, j_2\})} \frac{\mu_{j_3}}{\eta_{j_3}(\{j_1, j_2, j_3\})} \cdots \frac{\mu_{j_m}}{\eta_{j_m}(S)}\,. \tag{4}$$

This of course only makes sense if it is independent of the order in which we put the servers in S, so it has to hold equally for all permutations of $j_1, \ldots, j_m$. In particular, for every S and $j, k \in S$ we obtain the recursion:

$$\frac{\eta_j(S)}{\eta_k(S)} = \frac{\eta_j(S \backslash \{k\})}{\eta_k(S \backslash \{j\})}. \tag{5}$$

When the system is in state S we denote by $\eta(S)$ the rate at which one of the idle servers will become busy. We get two expressions for $\eta(S)$: it is the sum of the $\eta_j(S)$, and it is the sum of the arrival rates of all the customer types which can be served by the servers in S:

$$\eta(S) = \sum_{j \in S} \eta_j(S) = \sum_{i \in C(S)} \lambda_i \tag{6}$$

Proposition 1 *The equations (5), (6) uniquely determine the values of $\eta_j(S)$ for all S and $j \in S$.*

Proof. For singletons $S = \{j\}$,

$$\eta_j(\{j\}) = \sum_{i \in C(j)} \lambda_i$$

We proceed by induction, assuming we have determined the unique values for all states S of size $m-1$. Consider then the state $S = \{j_1, \dots, j_m\}$, and a server $k \in S$. From (5) and the induction hypothesis we obtain:

$$\frac{\eta(S)}{\eta_k(S)} = \frac{\eta_{j_1}(S) + \cdots + \eta_{j_m}(S)}{\eta_k(S)} = 1 + \sum_{j \in S \backslash \{k\}} \frac{\eta_j(S \backslash \{k\})}{\eta_k(S \backslash \{j\})},$$

where $\eta(S)$ is also known, from (6). Hence:

$$\eta_k(S) = \eta(S) \Big/ \Big(1 + \sum_{j \in S \backslash \{k\}} \frac{\eta_j(S \backslash \{k\})}{\eta_k(S \backslash \{j\})}\Big). \tag{7}$$

□

Example 1, continued:

We calculate the $\eta_j(S)$, from the values of λ_1, λ_2:

$$\eta(\{1\}) = \eta_1(\{1\}) = \lambda_1 + \lambda_2, \qquad \eta(\{2\}) = \eta_2(\{2\}) = \lambda_1, \qquad \eta(\{1,2\}) = \lambda_1 + \lambda_2,$$

and using the recursion step:

$$\eta_1(\{1,2\}) = \eta(\{1,2\}) \Big/ \Big(1 + \frac{\eta_2(\{2\})}{\eta_1(\{1\})}\Big) = (\lambda_1 + \lambda_2) \Big/ \Big(1 + \frac{\lambda_1}{\lambda_1 + \lambda_2}\Big) = \frac{(\lambda_1 + \lambda_2)^2}{2\lambda_1 + \lambda_2},$$

$$\eta_2(\{1,2\}) = \eta(\{1,2\}) \Big/ \Big(1 + \frac{\eta_1(\{1\})}{\eta_2(\{2\})}\Big) = (\lambda_1 + \lambda_2) \Big/ \Big(1 + \frac{\lambda_1 + \lambda_2}{\lambda_1}\Big) = \frac{\lambda_1(\lambda_1 + \lambda_2)}{2\lambda_1 + \lambda_2}.$$

The stationary probabilities (2) follow now from (4).

3 Assigning customers to servers

In this section we show that it is possible to choose the assigning probabilities $P(i, j|S)$ so that the resulting $X(t)$ will be reversible, with transition rates and stationary distribution as determined in Section 2.

Having calculated the values $\eta_j(S)$ we now look for the assignment probabilities $P(i, j|S)$ so that

$$\eta_j(S) = \sum_{i \in C(j)} \lambda_i P(i, j|S). \tag{8}$$

Proposition 2 *There exist assignment probabilities $P(i, j|S)$, for all S, $j \in S$, and $i \in C(j)$, which satisfy (8).*

We prove Proposition 2 in four steps. The first one is the translation to a maximal flow problem [9].

Proposition 3 *To satisfy (8) for S we need to solve a maximal flow problem.*

Proof. Summing over all the servers in S,

$$\eta(S) = \sum_{j \in S} \eta_j(S) = \sum_{j \in S} \sum_{i \in C(j)} \lambda_i P(i, j|S) = \sum_{i \in C(S)} \lambda_i.$$

We formulate a maximal flow problem with nodes a, b and nodes $j \in S$, $i \in C(S)$, where there is an arc with infinite capacity from i to j if $i \in C(j)$, and there are arcs from a to i with capacity λ_i and arcs from j to b with capacity $\eta_j(S)$ (see Fig 1).

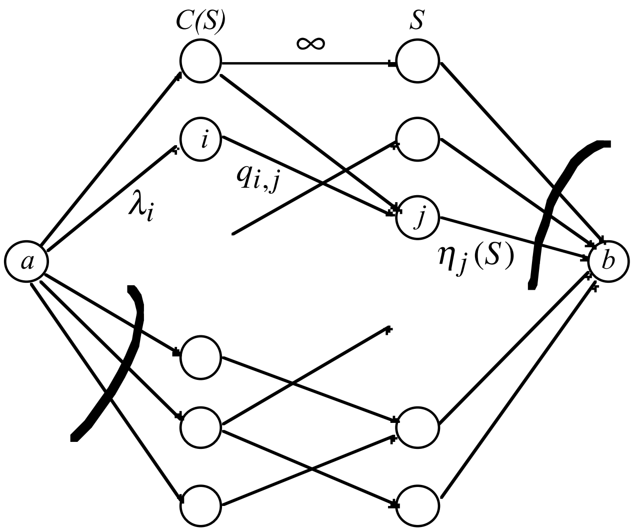

Figure 1: A maximal flow problem for finding $P(i, j|S)$

If the maximal flow in this network is $\eta(S)$, and $q_{i,j}$ is the flow on the arc from i to j, then $P(i, j|S) = q_{i,j}/\lambda_i$ solve (8). □

Example 1, concluded:

We calculate the assignment probabilities by solving:

$$\eta_2(\{1, 2\}) = \lambda_1 P(1, 2|\{1, 2\}), \qquad \eta_1(\{1, 2\}) = \lambda_2 + \lambda_1 P(1, 1|\{1, 2\}),$$

to obtain the values (1).

Proposition 4 *A necessary and sufficient condition for the existence of a flow of $\eta(S)$ in the network is: for every $R \subseteq S$*

$$\sum_{i \in C(R)} \lambda_i \geq \sum_{j \in R} \eta_j(S). \tag{9}$$

Proof. See the proof of Proposition 4 in [8]. □

Proposition 5 *A sufficient condition for (9) is that η satisfy the following monotonicity condition: for all $j \in R \subseteq S$*

$$\eta_j(R) \geq \eta_j(S). \tag{10}$$

Proof. Note that (9) actually says:

$$\sum_{j\in R}\eta_j(R)=\eta(R)=\sum_{i\in C(R)}\lambda_i\geq\sum_{j\in R}\eta_j(S).$$

which is clearly implied by (10). $\square$

Proposition 6 *The monotonicity condition (10) always holds.*

Proof. The proof is by induction on the size of S, the case $R=S$ (in particular $R=S=\{j\}$) is trivial. It is enough to verify the condition for R and S differing by only one element, say $S=R\cup\{q\}$. Suppose S has two or more elements and monotonicity has been established for smaller sets.

Then, for $k\in R$ ($k\neq q$), by (7),

$$\eta_k(R)=\eta(R)\Big/\Big(1+\sum_{j\in R\backslash\{k\}}\frac{\eta_j(R\backslash\{k\})}{\eta_k(R\backslash\{j\})}\Big),$$

$$\eta_k(S)=\eta(S)\Big/\Big(1+\sum_{j\in R\backslash\{k\}}\frac{\eta_j(S\backslash\{k\})}{\eta_k(S\backslash\{j\})}+\frac{\eta_q(S\backslash\{k\})}{\eta_k(S\backslash\{q\})}\Big).$$

The monotonicity property $\eta_k(R)\geq\eta_k(S)$ can be rewritten as

$$\eta(S)\left[1+\sum_{j\in R\backslash\{k\}}\frac{\eta_j(R\backslash\{k\})}{\eta_k(R\backslash\{j\})}\right]\leq\eta(R)\left[1+\sum_{j\in R\backslash\{k\}}\frac{\eta_j(S\backslash\{k\})}{\eta_k(S\backslash\{j\})}+\frac{\eta_q(S\backslash\{k\})}{\eta_k(S\backslash\{q\})}\right].$$

We can rearrange the left hand side:

$$\eta(S)\left[1+\sum_{j\in R\backslash\{k\}}\frac{\eta_j(R\backslash\{k\})}{\eta_k(R\backslash\{j\})}\right]=\eta(R)\left[1+\sum_{j\in R\backslash\{k\}}\frac{\eta_j(R\backslash\{k\})}{\eta_k(R\backslash\{j\})}\right]+\big(\eta(S)-\eta(R)\big)\frac{\eta(R)}{\eta_k(R)}.$$

Hence we need to verify that

$$\left[1+\sum_{j\in R\backslash\{k\}}\frac{\eta_j(R\backslash\{k\})}{\eta_k(R\backslash\{j\})}\right]+\frac{\big(\eta(S)-\eta(R)\big)}{\eta_k(R)}\leq\left[1+\sum_{j\in R\backslash\{k\}}\frac{\eta_j(S\backslash\{k\})}{\eta_k(S\backslash\{j\})}+\frac{\eta_q(S\backslash\{k\})}{\eta_k(S\backslash\{q\})}\right].\qquad(11)$$

Note that for $j\in R\backslash\{k\}$:

$$\begin{aligned}\frac{\eta_j(R\backslash\{k\})}{\eta_k(R\backslash\{j\})}&=\frac{\eta_j(R\backslash\{k\})-\eta_j(S\backslash\{k\})}{\eta_k(R\backslash\{j\})}+\frac{\eta_j(S\backslash\{k\})}{\eta_k(R\backslash\{j\})}\\&\leq\frac{\eta_j(R\backslash\{k\})-\eta_j(S\backslash\{k\})}{\eta_k(R)}+\frac{\eta_j(S\backslash\{k\})}{\eta_k(S\backslash\{j\})},\end{aligned}$$

where the inequality follows by applying the induction hypothesis three times, yielding $\eta_j(R\backslash\{k\})-\eta_j(S\backslash\{k\})\geq 0$, $\eta_k(R\backslash\{j\})\geq\eta_k(R)$, and $\eta_k(R\backslash\{j\})\geq\eta_k(S\backslash\{j\})$. Taking the sum over $j\in R\backslash\{k\}$ yields

$$\begin{aligned}\sum_{j\in R\backslash\{k\}}\frac{\eta_j(R\backslash\{k\})}{\eta_k(R\backslash\{j\})}&\leq\sum_{j\in R\backslash\{k\}}\frac{\eta_j(R\backslash\{k\})-\eta_j(S\backslash\{k\})}{\eta_k(R)}+\sum_{j\in R\backslash\{k\}}\frac{\eta_j(S\backslash\{k\})}{\eta_k(S\backslash\{j\})}\\&=\frac{\eta(R\backslash\{k\})-\eta(S\backslash\{k\})+\eta_q(S\backslash\{k\})}{\eta_k(R)}+\sum_{j\in R\backslash\{k\}}\frac{\eta_j(S\backslash\{k\})}{\eta_k(S\backslash\{j\})}.\end{aligned}\qquad(12)$$

Finally we note that

$$\eta(S) - \eta(R) \leq \eta(S\backslash\{k\}) - \eta(R\backslash\{k\}), \tag{13}$$

since the left hand side is the sum of arrival rates over $C(q) \cap (\mathcal{C}\backslash C(R))$, while the right hand side is the sum of arrival rates of the larger or equal set $C(q) \cap (\mathcal{C}\backslash C(R\backslash\{k\}))$.

Combining (12) and (13) proves (11). $\square$

Example 2:

There are three customer types and three servers, with $C(1) = \{2,3\}$, $C(2) = \{1,3\}$, $C(3) = \{1,2\}$. Let $\lambda = \lambda_1 + \lambda_2 + \lambda_3$. For $i \neq j \neq k$ (note the symmetry) we have:

$$\eta_i(\{i\}) = \lambda_j + \lambda_k, \qquad \eta_i(\{i,j\}) = \frac{\lambda(\lambda_j + \lambda_k)}{\lambda_i + \lambda_j + 2\lambda_k},$$

$$\eta_i(\{i,j,k\}) = \frac{\lambda(\lambda^2 - \lambda_i^2)}{3\lambda^2 - \lambda_1^2 - \lambda_2^2 - \lambda_3^2},$$

and hence:

$$\pi(\{i\}) = \pi(\emptyset)\frac{\mu_i}{\lambda_j + \lambda_k}, \qquad \pi(\{i,j\}) = \pi(\emptyset)\frac{\mu_i\mu_j(\lambda_i + \lambda_j + 2\lambda_k)}{\lambda(\lambda_i + \lambda_k)(\lambda_j + \lambda_k)},$$

$$\pi(\{i,j,k\}) = \pi(\emptyset)\frac{\mu_i\mu_j\mu_k(3\lambda^2 - \lambda_1^2 - \lambda_2^2 - \lambda_3^2)}{\lambda^2(\lambda_i + \lambda_j)(\lambda_i + \lambda_k)(\lambda_j + \lambda_k)}.$$

We now look for the assignment probabilities. We get immediately for S of one or two servers:

$$P(j,i|\{i\}) = P(k,i|\{i\}) = 1, \qquad P(i,j|\{i,j\} = P(j,i|\{i,j\} = 1,$$

$$P(k,i|\{i,j\}) = \frac{\lambda_i + \lambda_k}{\lambda_i + \lambda_j + 2\lambda_k}, \qquad P(k,j|\{i,j\}) = \frac{\lambda_j + \lambda_k}{\lambda_i + \lambda_j + 2\lambda_k}.$$

When all three servers are idle, the equations to be solved are the three equations of the form:

$$\lambda_j P(j,i|\{1,2,3\}) + \lambda_k P(k,i|\{1,2,3\}) = \frac{\lambda(\lambda^2 - \lambda_i^2)}{3\lambda^2 - \lambda_1^2 - \lambda_2^2 - \lambda_3^2} \tag{14}$$

As we have seen, these equations do have positive solutions, but here there are three unknowns and only two equations (the three equations are dependent), so the solution is not unique. Using the abbreviations $P(i,j) \equiv P(i,j|\{1,2,3\})$ and $\eta_j \equiv \eta_j(\{1,2,3\})$, the solutions to (14) can be parameterized as:

$$\begin{bmatrix} P(i,j) \\ P(j,k) \\ P(k,i) \end{bmatrix} = \begin{bmatrix} 1 - P(i,k) \\ 1 - P(j,i) \\ 1 - P(k,j) \end{bmatrix} = (1-\theta)\begin{bmatrix} \frac{\max(0,\eta_j - \lambda_k, \lambda_i - \eta_k)}{\lambda_i} \\ \frac{\max(0,\eta_k - \lambda_i, \lambda_j - \eta_i)}{\lambda_j} \\ \frac{\max(0,\eta_i - \lambda_j, \lambda_k - \eta_j)}{\lambda_k} \end{bmatrix} + \theta \begin{bmatrix} \frac{\min(\lambda_i, \eta_j, \lambda_i + \lambda_j - \eta_k)}{\lambda_i} \\ \frac{\min(\lambda_j, \eta_k, \lambda_j + \lambda_k - \eta_i)}{\lambda_j} \\ \frac{\min(\lambda_k, \eta_i, \lambda_i + \lambda_k - \eta_j)}{\lambda_k} \end{bmatrix}$$

where $0 \leq \theta \leq 1$.

Example 2 illustrates two important points. First, the assignment probabilities need not be unique. Second, one can ask: Is it true that $P(i,j|S_1) = P(i,j|S_2)$ if $S(i) \cap S_1 = S(i) \cap S_2$? In other words, given the set of idle servers which can serve i, the $P(i,j|\cdot)$ do not depend on additional available servers which cannot serve i. This is false, as Example 2 shows: If we take $P(i,k|\{1,2,3\}) = P(i,k|\{j,k\})$, $P(j,k|\{1,2,3\}) = P(j,k|\{i,k\})$ and $P(k,i|\{1,2,3\}) = P(k,i|\{i,j\})$, this choice will not satisfy the equations (14). So, if we want to have a product form solution, the routing rates have to change every time the state changes, even if the routing options for some of the customer types do not change. This shows how fragile the phenomenon of product form is.

4 Discussion

In this paper we considered a loss system. It is interesting to also investigate the same system with no losses: this is a single queueing station, with multi-type customers queueing in I different queues, and J servers, which are heterogeneous, server j serving the queues of customers of types $C(j)$, at rate μ_j.

In that case one needs to specify the service policy. A very common service policy is FCFS: whenever a server becomes available he will serve the longest waiting customer which is compatible with him, or else he will idle. One needs also to specify assignment rules for customers which arrive and find suitable servers which are idle. It again turns out that the assignment probabilities can be chosen so that this FCFS system will satisfy partial balance, and have a product form stationary distribution. This topic was explored in [1, 2, 16, 3], and finally resolved in [14, 15].

It is also hoped that these results will help in the analysis of some models and open problems discussed in [13, 8, 4].

Acknowledgement

The authors would like to thank Scott Provan for many stimulating discussions on this problem.

References

[1] Adan, I.J.B.F., Wessels, J., Zijm, W.H.M. (1989) Queuing analysis in a flexible assembly system with a job-dependent parallel structure. *Operations Research Proceedings 1988* Springer-Verlag, Berlin, pp. 551–558.

[2] Adan, I.J.B.F., Wessels, J., Zijm, W.H.M. (1991) Flexible assembly and shortest queue problems. *Modern production concepts, theory and applications*, G. Fandel, G. Zaepfel (eds.), Springer-Verlag, Berlin, pp. 644–659.

[3] Adan, I., Foley, R., McDonald, D. (2009) Exact Asymptotics of the Stationary Distribution of a Markov Chain: a Production Model. *Queueing Systems* 62: 311–344.

[4] Adan, I.J.B.F., Caldentey, R., Kaplan, E.H., Weiss, G., (2010) First come first served matching of multi type customers and multi type servers. Preprint.

[5] Aerts, J., Korst, J., Verhaegh, W., (2007) Load balancing for redundant storage strategies: Multiprocessor scheduling with machine eligibility. *Journal of Scheduling* 4: 245–257.

[6] Aksin, Z., Armony, M., Mehrotra, V., (2007) The modern call-center, a multi-disciplinary perspective on operations management research. *Production and Operations Management* Special Issue on Service Operations in honor of John Buzacott (ed. G. Shanthikumar and D. Yao) 16 (6): 655–688.

[7] de Bruin, A.M., Bekker, R., van Zanten, L., Koole, G.M. (2009) Dimensioning hospital wards using the Erlang loss model. Annals of Operations Research, to appear.

[8] Caldentey, R., Kaplan, E.H., Weiss, G., (2009) FCFS infinite bipartite matching of servers and customers. *Advances in Applied Probability* 41: 695–730.

[9] Ford, L. R.,Jr., Fulkerson, D. R. (1962) *Flows in Networks*, Princeton University Press, Princeton.

[10] Garnett, O., Mandelbaum, A. (2000) An Introduction to Skill-Based Routing and its Operational Complexities. http://iew3.technion.ac.il/serveng/Lectures/SBR.pdf

[11] Kelly, F.P. (1979) *Reversibility and Stochastic Networks*, Wiley.

[12] McManus, M.L., Long, M.C., Copper, A., Litvak, E. (2004) Queueing theory accurately models the need for critical care resources. *Anesthesiology* 100: 1271-1276.

[13] Talreja, R., Whitt, W. (2007) Fluid Models for Overloaded Multi-class many-service queueing systems with FCFS routing. *Management Science* 54: 1513–1527.

[14] Visschers, J.W.C.H. (2000) *Random walks with geometric jumps*. Ph.D. Thesis, Eindhoven University of Technology.

[15] Visschers, J.W.C.H., Adan, I.J.B.F., Weiss, G.(2009). A multi-class multi-server system with customer-class service restrictions. In preparation.

[16] Zijm, W.H.M., Laarhoven, P.J.M. (1993) Production Preparation and Numerical Control in PCB assembly. *International Journal of Flexible Manufacturing Systems* 5(3): 187–207.