EURANDOM PREPRINT SERIES
2011-042 September 12, 2011

A loss system with skill based servers under assign to longest idle server policy

I. Adan, G. Weiss
ISSN 1389-2355

A loss system with skill based servers under assign to longest idle server policy

Ivo Adan[*] Gideon Weiss[†]

September 12, 2011

Abstract

We consider a memoryless loss system with servers $\mathcal{S} = \{1, \dots, J\}$, and with customer types $\mathcal{C} = \{1, \dots, I\}$. Servers are multi-type: server j works at rate μ_j, and can serve a subset of customer types $C(j)$. An arriving customer will go to the longest idling server which can serve him, or be lost. We obtain a simple explicit steady state distribution for this system, and calculate various performance measures of this system in steady state. We provide some illustrative examples. We compare this system with a similar system discussed recently by Adan, Hurkens and Weiss [2]. We also show that this system is insensitive, the results hold also for general processing time distributions.

Keywords: Service system; loss system; multi type customers; multi type servers; product form solution; go to longest waiting server policy; insensitivity.

1 Model

We consider a loss system with servers $\mathcal{S} = \{1, \dots, J\}$, and with customer types $\mathcal{C} = \{1, \dots, I\}$. Arrivals are Poisson. Customers of type i arrive at rate λ_i. Service is skill based, so that server j can serve a subset of customer types $C(j)$. The service times of server j are independent and exponentially distributed with mean $1/\mu_j$. We use the following notation: for a subset of servers S we let $\mu_S = \sum_{j \in S} \mu_j$ be the combined service rate of servers in S, and we let $C(S) = \bigcup_{j \in S} C(j)$ be the set of customer types that can be served by one or more servers in S. For a set of customer types C we define λ_C and $S(C)$ similarly, i.e., λ_C is the combined arrival rate of customers in C and $S(C)$ is the set of servers that can serve one or more customer types in C.

The system is a loss system: customers that arrive, and do not find an idle server which can serve them, are lost. Customers that find more than one server that can serve them, go to the server that has been idle for the longest time. We call this a loss system with assign to the longest idle server (ALIS) regime.

We define the state of the system at time t as $X(t) = s$, where $s = (j_1, j_2, \dots, j_m)$ is the list of idle servers at time t, ordered by their order of becoming idle, so that server j_1 has been idle for the longest time, and so on. With this state definition, under the assign to the longest idle server regime, $X(t)$ is a continuous time finite state Markov chain (CTMC).

[*]Department of Mechanical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, the Netherlands; email iadan@tue.nl Research supported in part by the Netherlands Organization for Scientific Research (NWO).

[†]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 Grant 711/09.

We derive the stationary distribution of this system, given by:

$$\pi(j_1,\ldots,j_m) = \pi(\emptyset)\,\frac{\mu_{j_1}}{\lambda_{\mathcal{C}(j_1)}}\,\frac{\mu_{j_2}}{\lambda_{\mathcal{C}(\{j_1,j_2\})}}\,\cdots\,\frac{\mu_{j_m}}{\lambda_{\mathcal{C}(\{j_1,\ldots,j_m\})}}. \tag{1}$$

in Section 2. We then derive expressions for various performance measures of the system in Section 3. In Section 4 we give some examples. These illustrate the effect of different skill levels of the servers, and the degree of overlap in skills. A similar system was studied in [2]. In that system arriving customers choose servers according to some special assignment probability distributions that depend on the type of the arrival and on the set of idle servers, but not on the order in which servers became idle. Surprisingly, we show in Section 5 that the stationary distribution of the set of idle servers is the same in both models. Finally, in Section 6 we show that these results are insensitive to the service time distributions.

This paper was motivated by the recent interest in service systems which provide skill based service. Such systems are common in communication and internet applications, in call centers, and in health systems. Our results here are part of a project in which we study the behavior of service systems with several types of customers and several types of servers, where a bipartite graph describes compatibility of server with customer type. In these skill based service systems customers can obtain specialized service, while the overlap in server skills may provide for server pooling and more efficient utilization of servers. In the present paper we consider loss systems, under a policy of assigning customers to the longest idle server. Related results for similar systems under ALIS and FCFS are presented in [6, 4, 1, 3, 7].

2 Partial balance and the stationary distribution

The equilibrium equations are, for an ordered list of idle servers $(j_1,\ldots,j_m)$, $1 \le m \le J$,

$$\begin{aligned} \pi(j_1,\ldots,j_m)\Big(\mu_{\{k:k\neq j_1,\ldots,j_m\}} + \lambda_{\mathcal{C}(\{j_1,\ldots,j_m\})}\Big) &= \pi(j_1,\ldots,j_{m-1})\mu_{j_m} \\ + \sum_{k\neq j_1,\ldots,j_m} \Big(\pi(k,j_1,\ldots,j_m)\lambda_{\mathcal{C}(k)} &+ \pi(j_1,k,\ldots,j_m)\lambda_{\mathcal{C}(k)\setminus\mathcal{C}(j_1)} + \cdots \\ &+\pi(j_1,\ldots,j_m,k)\lambda_{\mathcal{C}(k)\setminus\mathcal{C}(\{j_1,\ldots,j_m\})}\Big). \end{aligned} \tag{2}$$

We show that they hold for π given by (1), and in fact partial balance holds, i.e.,

$$\pi(j_1,\ldots,j_m)\lambda_{\mathcal{C}(\{j_1,\ldots,j_m\})} = \pi(j_1,\ldots,j_{m-1})\mu_{j_m} \tag{3}$$

and for all $k \neq j_1,\ldots,j_m$:

$$\begin{aligned} \pi(j_1,\ldots,j_m)\mu_k = \Big(&\pi(k,j_1,\ldots,j_m)\lambda_{\mathcal{C}(k)} + \pi(j_1,k,\ldots,j_m)\lambda_{\mathcal{C}(k)\setminus\mathcal{C}(j_1)} + \cdots \\ &+\pi(j_1,\ldots,j_m,k)\lambda_{\mathcal{C}(k)\setminus\mathcal{C}(\{j_1,\ldots,j_m\})}\Big). \end{aligned} \tag{4}$$

Theorem 1 *The stationary distribution π is given by (1), and it obeys the partial balance equations (3) and (4).*

Proof. The verification of (3) is immediate: we substitute (1) to get

$$\pi(j_1,\ldots,j_m)\lambda_{\mathcal{C}(\{j_1,\ldots,j_m\})}$$

$$
\begin{aligned}
&= \pi(\emptyset)\ \frac{\mu_{j_1}}{\lambda_{\mathcal{C}(j_1)}}\ \frac{\mu_{j_2}}{\lambda_{\mathcal{C}(\{j_1,j_2\})}}\ \cdots\ \frac{\mu_{j_m}}{\lambda_{\mathcal{C}(\{j_1,\dots,j_m\})}} \times \lambda_{\mathcal{C}(\{j_1,\dots,j_m\})} \\
&= \pi(\emptyset)\ \frac{\mu_{j_1}}{\lambda_{\mathcal{C}(j_1)}}\ \frac{\mu_{j_2}}{\lambda_{\mathcal{C}(\{j_1,j_2\})}}\ \cdots\ \frac{\mu_{j_{m-1}}}{\lambda_{\mathcal{C}(\{j_1,\dots,j_{m-1}\})}} \times \mu_{j_m} \\
&= \pi(j_1,\dots,j_{m-1})\mu_{j_m}.
\end{aligned}
$$

To show (4) we first show that (1) satisfies the equation

$$
\begin{aligned}
&\pi(k,j_1,\dots,j_m)\lambda_{\mathcal{C}(k)} + \cdots + \pi(j_1,\dots,j_l,k,j_{l+1}\dots,j_m)\lambda_{\mathcal{C}(k)\setminus\mathcal{C}(\{j_1,\dots,j_l\})} \\
&= \pi(j_1,\dots,j_l,k,j_{l+1}\dots,j_m)\lambda_{\mathcal{C}(\{k,j_1,\dots,j_l\})}.
\end{aligned} \tag{5}
$$

We prove (5) by induction on l. There is nothing to prove for $l=0$. Assuming (5) holds for l we get for $l+1$:

$$
\begin{aligned}
&\pi(k,j_1,\dots,j_m)\lambda_{\mathcal{C}(k)} + \cdots + \pi(j_1,\dots,j_l,j_{l+1},k,\dots,j_m)\lambda_{\mathcal{C}(k)\setminus\mathcal{C}(\{j_1,\dots,j_{l+1}\})} \\
&= \pi(j_1,\dots,j_l,k,j_{l+1}\dots,j_m)\lambda_{\mathcal{C}(\{k,j_1,\dots,j_l\})} + \pi(j_1,\dots,j_l,j_{l+1},k,\dots,j_m)\lambda_{\mathcal{C}(k)\setminus\mathcal{C}(\{j_1,\dots,j_{l+1}\})} \\
&= \pi(\emptyset)\ \frac{\mu_{j_1}\cdots\mu_{j_m}\mu_k}{\lambda_{\mathcal{C}(j_1)}\cdots\lambda_{\mathcal{C}(\{j_1,\dots,j_l\})}\lambda_{\mathcal{C}(\{k,j_1,\dots,j_l\})}\ \lambda_{\mathcal{C}(\{k,j_1,\dots,j_l,j_{l+1}\})}\cdots\lambda_{\mathcal{C}(\{k,j_1,\dots,j_m\})}} \times \lambda_{\mathcal{C}(\{k,j_1,\dots,j_l\})} \\
&\quad +\pi(\emptyset)\ \frac{\mu_{j_1}\cdots\mu_{j_m}\mu_k}{\lambda_{\mathcal{C}(j_1)}\cdots\lambda_{\mathcal{C}(\{j_1,\dots,j_l\})}\lambda_{\mathcal{C}(\{j_1,\dots,j_l,j_{l+1}\})}\ \lambda_{\mathcal{C}(\{k,j_1,\dots,j_l,j_{l+1}\})}\cdots\lambda_{\mathcal{C}(\{k,j_1,\dots,j_m\})}} \times \lambda_{\mathcal{C}(k)\setminus\mathcal{C}(\{j_1,\dots,j_{l+1}\})} \\
&= \pi(\emptyset)\ \frac{\mu_{j_1}\cdots\mu_{j_m}\mu_k}{\lambda_{\mathcal{C}(j_1)}\cdots\lambda_{\mathcal{C}(\{j_1,\dots,j_l\})}\quad \lambda_{\mathcal{C}(\{k,j_1,\dots,j_l,j_{l+1}\})}\cdots\lambda_{\mathcal{C}(\{k,j_1,\dots,j_m\})}} \left[1+\frac{\lambda_{\mathcal{C}(k)\setminus\mathcal{C}(\{j_1,\dots,j_{l+1}\})}}{\lambda_{\mathcal{C}(\{j_1,\dots,j_l,j_{l+1}\})}}\right] \\
&= \pi(\emptyset)\ \frac{\mu_{j_1}\cdots\mu_{j_m}\mu_k}{\lambda_{\mathcal{C}(j_1)}\cdots\lambda_{\mathcal{C}(\{j_1,\dots,j_l\})}\lambda_{\mathcal{C}(\{j_1,\dots,j_l,j_{l+1}\})}\lambda_{\mathcal{C}(\{k,j_1,\dots,j_l,j_{l+1}\})}\cdots\lambda_{\mathcal{C}(\{k,j_1,\dots,j_m\})}} \times \lambda_{\mathcal{C}(\{k,j_1,\dots,j_{l+1}\})} \\
&= \pi(j_1,\dots,j_l,j_{l+1},k,\dots,j_m) \times \lambda_{\mathcal{C}(\{k,j_1,\dots,j_{l+1}\})},
\end{aligned}
$$

which is what we needed to show for (5).

We now get that the right hand side of (4), when substituting (1) and using (5) for $l=m$ equals:

$$
\begin{aligned}
&\pi(k,j_1,\dots,j_m)\lambda_{\mathcal{C}(k)} + \pi(j_1,k,\dots,j_m)\lambda_{\mathcal{C}(k)\setminus\mathcal{C}(j_1)} + \cdots + \pi(j_1,\dots,j_m,k)\lambda_{\mathcal{C}(k)\setminus\mathcal{C}(\{j_1,\dots,j_m\})} \\
&\quad = \pi(j_1,\dots,j_m,k)\lambda_{\mathcal{C}(\{k,j_1,\dots,j_m\})} \\
&\quad = \pi(\emptyset)\ \frac{\mu_{j_1}}{\lambda_{\mathcal{C}(j_1)}}\ \frac{\mu_{j_2}}{\lambda_{\mathcal{C}(\{j_1,j_2\})}}\ \cdots\ \frac{\mu_{j_m}}{\lambda_{\mathcal{C}(\{j_1,\dots,j_m\})}} \frac{\mu_k}{\lambda_{\mathcal{C}(k,\{j_1,\dots,j_m\})}} \times \lambda_{\mathcal{C}(\{k,j_1,\dots,j_m\})} \\
&\quad = \pi(\emptyset)\ \frac{\mu_{j_1}}{\lambda_{\mathcal{C}(j_1)}}\ \frac{\mu_{j_2}}{\lambda_{\mathcal{C}(\{j_1,j_2\})}}\ \cdots\ \frac{\mu_{j_m}}{\lambda_{\mathcal{C}(\{j_1,\dots,j_m\})}}\mu_k,
\end{aligned}
$$

which is exactly the left hand side of (4), when substituting (1). This completes the proof. □

3 Calculation of various performance measures

In this section we use the steady state distribution of the system to calculate various performance measures. These include the fraction of time each server is busy, the fraction of customers of each type that are lost, and the rate at which customers of type i are served by servers of type j.

3.1 The fraction of busy time of each server

Let ν_j be the fraction of time that server j is busy. Then:

$$\nu_j = \sum_{(j_1,\ldots,j_m):j\notin\{j_1,\ldots,j_m\}} \pi(j_1,\ldots,j_m),$$

i.e., we sum up all the steady state probabilities of all the states in which j is not idle.

For a subset of servers S denote by

$$B_S = 1\Big/ \sum_{(j_1,\ldots,j_m):\{j_1,\ldots,j_m\}\subset S} \frac{\mu_{j_1}}{\lambda_{\mathcal{C}(j_1)}} \frac{\mu_{j_2}}{\lambda_{\mathcal{C}(\{j_1,j_2\})}} \cdots \frac{\mu_{j_m}}{\lambda_{\mathcal{C}(\{j_1,\ldots,j_m\})}}.$$

Clearly

$$\pi(\emptyset) = B_{\{1,\ldots,J\}} = 1\Big/ \sum_{\text{all } (j_1,\ldots,j_m)} \frac{\mu_{j_1}}{\lambda_{\mathcal{C}(j_1)}} \frac{\mu_{j_2}}{\lambda_{\mathcal{C}(\{j_1,j_2\})}} \cdots \frac{\mu_{j_m}}{\lambda_{\mathcal{C}(\{j_1,\ldots,j_m\})}}$$

and we see that

$$\nu_j = \pi(\emptyset) \sum_{(j_1,\ldots,j_m):j\notin\{j_1,\ldots,j_m\}} \frac{\mu_{j_1}}{\lambda_{\mathcal{C}(j_1)}} \frac{\mu_{j_2}}{\lambda_{\mathcal{C}(\{j_1,j_2\})}} \cdots \frac{\mu_{j_m}}{\lambda_{\mathcal{C}(\{j_1,\ldots,j_m\})}} = B_{\{1,\ldots,J\}}/B_{\{1,\ldots,J\}\setminus\{j\}}.$$

3.2 The fraction of customers of each type that are lost

Let θ_i be the fraction of customers of type i which are lost.

Arrivals occur as a Poisson stream, hence an arrival of type i sees the system in its time average steady state. A customer of type i will be lost if the set of idle servers does not contain any server from $S(i)$, hence, similar to the derivation of ν_j:

$$\theta_i = \sum_{(j_1,\ldots,j_m):S(i)\cap\{j_1,\ldots,j_m\}=\emptyset} \pi(j_1,\ldots,j_m) = B_{\{1,\ldots,J\}}/B_{\{1,\ldots,J\}\setminus S(i)}.$$

3.3 The matching rate of server customer pairs

Let $r_{i,j}$ be the rate at which customer of type i are matched with servers of type j. A customer of type i arrives at rate λ_i, and will be matched with the first (longest idle) server available with which it is compatible. The calculation of $r_{i,j}/\lambda_i$ will then consist of summation over of all the steady state probabilities of states in which an arriving customer of type i is matched to a server of type j, where $j \in S(i)$. This summation includes first all the states $(i_1,\ldots,i_m)$ in which $i_1 = j$, second all the states in which $i_1 \notin S(i)$ and $i_2 = j$, and so on, where for any $k \le J$ the summation is over all states in which there are at least k idle servers, and $\{i_1,\ldots,i_{k-1}\}\cap S(i) = \emptyset$ while $i_k = j$.

For state $s = (j_1,\ldots,j_m)$ we define $1_{i\to j|s}$ as the indicator that an i,j match occurs in state s, and this is the sum of the indicators:

$$1_{i\to j|s} = 1_{j_1=j} + 1_{j_1\notin S(i) \text{ and } j_2=j} + \cdots + 1_{\{j_1,\ldots,j_{m-1}\}\cap S(i)=\emptyset \text{ and } j_m=j}$$

and we then have:

$$r_{i,j} = \lambda_i \sum_s \pi(s) 1_{i\to j|s}.$$

4 Examples

The calculation of the steady state probabilities and of the additional performance measures involves summation over all permutations of all subsets, which is quite laborious. It may well be that this computation is $\sharp P$-complete. We have programmed these calculations, and are able to perform them for reasonably small J. We consider here three examples. The first is an example of a network with servers of three different skill levels. The second example is a symmetric system where each server serves two types of customers. The third example extends the second example, by considering symmetric systems with varying degrees of overlap in the server skills, and the resulting performance for the different degrees of overlap is compared.

4.1 Example 1: An example of servers with different skill levels

We consider a system with 3 types of customers and 3 servers. The servers have different skill levels, with server 1 serving only customers of type 1, server 2 serving customers of types 1 and 2, and server 3 serving all types of customers, i.e. $C(1) = \{1\}$, $C(2) = \{1, 2\}$, $C(3) = \{1, 2, 3\}$, as in Figure 1. The arrival rates of the 3 types are $\lambda_1 = 2$, $\lambda_2 = 1$, $\lambda_3 = 1$ and the service rates are $\mu_1 = 1$, $\mu_2 = 1$, $\mu_3 = 2$.

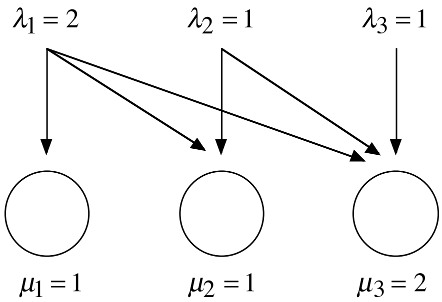


Figure 1: A system with 3 servers of different skill levels

There is a total of 16 states. For the different states of the system we get the following steady state probabilities:

$$\begin{array}{cccccc}
 & & \pi(\emptyset) = \frac{12}{43}, & & & \\
 & \pi(1) = \frac{6}{43}, & \pi(2) = \frac{4}{43}, & \pi(3) = \frac{6}{43}, & & \\
\pi(1,2) = \frac{2}{43}, & \pi(1,3) = \frac{3}{43}, & \pi(2,1) = \frac{4}{129}, & \pi(2,3) = \frac{2}{43}, & \pi(3,1) = \frac{3}{86}, & \pi(3,2) = \frac{3}{86}, \\
\pi(1,2,3) = \frac{1}{43}, & \pi(1,3,2) = \frac{3}{172}, & \pi(2,1,3) = \frac{2}{129}, & \pi(2,3,1) = \frac{1}{86}, & \pi(3,1,2) = \frac{3}{344}, & \pi(3,2,2) = \frac{3}{344}.
\end{array}$$

The matching rates for customer server pairs are given in the following table:

$i \ \backslash \ j$	1	2	3	Lost
1	0.593	0.395	0.454	0.558
2	—	0.267	0.314	0.419
3	—	—	0.411	0.589
Idle	0.407	0.337	0.821	

Table 1: Matching rates $r_{i,j}$ for the example in Figure 1

In the table above, the column for loss is the rate at which calls are lost, and the row for idle is the lost service rate due to idleness. Rows add up to λ_i, columns to μ_j.

4.2 Example 2: A symmetric system with small overlap of skills

We consider a system with 5 types of customers and 5 servers, where each server can serve 2 types of customers, $C(j) = \{j, j+1\}$ where in this example all indices are taken modulo 5, as shown in Figure 2.

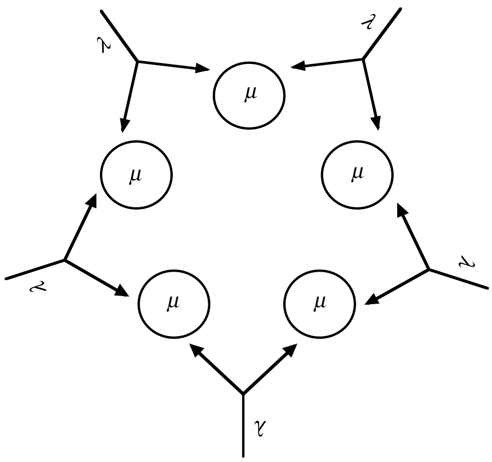


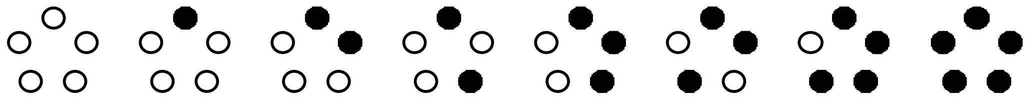
Figure 2: A symmetric system with 5 types of customers and 5 servers, with small overlap

The total number of states is $J! \sum_{k=0}^{J} \frac{1}{k!} = 326$, but by the symmetry we need to consider only a few configurations of idle servers as shown in Figure 3.

The calculation of steady state probabilities is summarized in Table 2. Note that the configurations in Figure 3 may correspond to multiple states, and states corresponding to the same configuration may have different steady state probabilities (such as states corresponding to (v)).

From this we get that

$$\pi(\emptyset) = 240 \Bigg/ \left(240 + 600\frac{\mu}{\lambda} + 700\left(\frac{\mu}{\lambda}\right)^2 + 475\left(\frac{\mu}{\lambda}\right)^3 + 190\left(\frac{\mu}{\lambda}\right)^4 + 38\left(\frac{\mu}{\lambda}\right)^5\right),$$



(*i*) (*ii*) (*iii*) (*iv*) (*v*) (*vi*) (*vii*) (*viii*)

Figure 3: Configurations of idle (black) and busy (white) servers

from which one can also read the steady state probabilities of the number of idle servers.

To calculate the probability that server j is busy, we use:

$$\nu = \nu_j = \frac{B_{\{1,2,3,4,5\}}}{B_{\{1,2,3,4\}}} = \frac{240 + 480\frac{\mu}{\lambda} + 420\left(\frac{\mu}{\lambda}\right)^2 + 190\left(\frac{\mu}{\lambda}\right)^3 + 38\left(\frac{\mu}{\lambda}\right)^4}{240 + 600\frac{\mu}{\lambda} + 700\left(\frac{\mu}{\lambda}\right)^2 + 475\left(\frac{\mu}{\lambda}\right)^3 + 190\left(\frac{\mu}{\lambda}\right)^4 + 38\left(\frac{\mu}{\lambda}\right)^5},$$

where the calculation of $B_{\{1,2,3,4\}}$ is similar to that of $B_{\{1,2,3,4,5\}}$.

The probability that a customer of type i will not obtain service is obtained similarly:

$$\theta = \theta_i = \frac{B_{\{1,2,3,4,5\}}}{B_{\{1,2,3\}}} = \frac{240 + 360\frac{\mu}{\lambda} + 220\left(\frac{\mu}{\lambda}\right)^2 + 55\left(\frac{\mu}{\lambda}\right)^3}{240 + 600\frac{\mu}{\lambda} + 700\left(\frac{\mu}{\lambda}\right)^2 + 475\left(\frac{\mu}{\lambda}\right)^3 + 190\left(\frac{\mu}{\lambda}\right)^4 + 38\left(\frac{\mu}{\lambda}\right)^5}.$$

By the symmetry, for i compatible with j:

$$r_{i,j} = \frac{1}{2}\lambda(1-\theta).$$

4.3 Example 3: Performance of symmetric system as influenced by level of overlap

It is interesting to compare the performance of our system for various levels of overlap of skills. We consider again the symmetric system of the last example, with 5 types of customers and 5 servers, and with equal arrival rates and equal service rates, but with different levels of overlap of skills. To be clear, we mean by the level of overlap the number of types that each server can serve. We then have overlap level ranging from 1 to 5, where for overlap level 1 each server serves only one type of customers and there is no overlap, while for overlap level of 5 each server serves all customer types for maximal overlap.

The system with overlap level 1 server j serves only customers of type j. Here the system uncouples into 5 single server loss systems.

The system with overlap level 5 each server serves all types and the system is the Erlang loss system with 5 servers.

The system with overlap level $L = 2,3,4$ server j serves customer types $j, j+1, \ldots, j+L-1$ (modulo 5), where we already discussed the symmetric system with overlap level 2.

The following table compares the performance of these systems. We include the following performance measures: The loss rate, fraction of customers lost of each type; the idleness rate, fraction of time that each server is idle; the distribution of the number of idle servers. The table lists these performance measures for three traffic intensities: light traffic, $\mu/\lambda = 2$, medium traffic, $\mu/\lambda = 1$, and heavy traffic, $\mu/\lambda = 0.5$.

The above table shows that performance strongly improves by increasing overlap, most prominently from no overlap to an overlap of 2. As expected, the loss rate is most strongly influenced by the level of overlap in light traffic, whereas the idle time is most strongly influenced by the level of overlap in heavy traffic.

No. of idle servers	Configuration	No. of states	Probability	Total
0	(i)	1	$\pi(\emptyset)$	$\pi(\emptyset)$
1	(ii)	5	$\pi(\emptyset)\frac{\mu}{\lambda}\frac{1}{2}$	$\pi(\emptyset)\frac{\mu}{\lambda}\frac{5}{2}$
2	(iii)	10	$\pi(\emptyset)\left(\frac{\mu}{\lambda}\right)^2\frac{1}{2}\frac{1}{3}$	$\pi(\emptyset)\left(\frac{\mu}{\lambda}\right)^2\frac{10}{6}$
2	(iv)	10	$\pi(\emptyset)\left(\frac{\mu}{\lambda}\right)^2\frac{1}{2}\frac{1}{4}$	$\pi(\emptyset)\left(\frac{\mu}{\lambda}\right)^2\frac{10}{8}$
3	(iii) then (v)	20	$\pi(\emptyset)\left(\frac{\mu}{\lambda}\right)^3\frac{1}{2}\frac{1}{3}\frac{1}{4}$	$\pi(\emptyset)\left(\frac{\mu}{\lambda}\right)^3\frac{20}{24}$
3	(iii) then (vi)	10	$\pi(\emptyset)\left(\frac{\mu}{\lambda}\right)^3\frac{1}{2}\frac{1}{3}\frac{1}{5}$	$\pi(\emptyset)\left(\frac{\mu}{\lambda}\right)^3\frac{10}{30}$
3	(iv) then (v)	10	$\pi(\emptyset)\left(\frac{\mu}{\lambda}\right)^3\frac{1}{2}\frac{1}{4}\frac{1}{4}$	$\pi(\emptyset)\left(\frac{\mu}{\lambda}\right)^3\frac{10}{32}$
3	(iv) then (vi)	20	$\pi(\emptyset)\left(\frac{\mu}{\lambda}\right)^3\frac{1}{2}\frac{1}{4}\frac{1}{5}$	$\pi(\emptyset)\left(\frac{\mu}{\lambda}\right)^3\frac{20}{40}$
4	(vii)	120	$\pi(j_1,j_2,j_3,j_4)=\pi(j_1,j_2,j_3)\frac{\mu}{\lambda}\frac{1}{5}$	$\pi(\emptyset)\left(\frac{\mu}{\lambda}\right)^4\frac{190}{240}$
5	(vii)	120	$\pi(j_1,j_2,j_3,j_4,j_5)=\pi(j_1,j_2,j_3,j_4)\frac{\mu}{\lambda}\frac{1}{5}$	$\pi(\emptyset)\left(\frac{\mu}{\lambda}\right)^5\frac{38}{240}$

Table 2: Calculation of steady state probabilities for the example in Figure 2

5 Comparison with a random assignment model

In [2] the same loss system is considered under a different regime. Instead of choosing the longest idle server, an arriving customer of type i which finds a set of idle servers S will choose server $j \in S$ with probability $P(i,j|S)$. This is different from our ALIS (assign to longest idle server) regime in two ways: first, the assignment does not depend on the order in which idle servers became available, and second, the assignment is random. Because only the set of idle servers is of interest in the system of [2], it can be described by a CTMC $Y(t)$ with state given by the set of idle servers S. It is shown in [2] that one can choose the assignment probabilities $P(i,j|S)$ so as to make $Y(t)$ reversible, in which case its steady state distribution is for every subset $S=\{j_1,\dots,j_m\}$ given by:

$$\pi_Y(\{j_1,\dots,j_m\}) = \pi_Y(\emptyset)\,\frac{\mu_{j_1}}{\eta_{j_1}(\{j_1\})}\,\frac{\mu_{j_2}}{\eta_{j_2}(\{j_1,j_2\})}\,\frac{\mu_{j_3}}{\eta_{j_3}(\{j_1,j_2,j_3\})}\,\cdots\,\frac{\mu_{j_m}}{\eta_{j_m}(S)}\,. \tag{6}$$

Here we use the notation

$$\eta(S) = \sum_{k\in S}\eta_k(S) = \lambda_{C(S)},$$

μ/λ	Overlap	Customer loss rate	Server idle time	Distribution of the number of idle servers 0	1	2	3	4	5
2	1	0.333	0.667	0.00412	0.0412	0.1646	0.3292	0.3292	0.1317
	2	0.185	0.593	0.0195	0.0976	0.228	0.309	0.347	0.0989
	3	0.117	0.559	0.0391	0.130	0.234	0.281	0.225	0.090
	4	0.0849	0.542	0.0566	0.141	0.226	0.272	0.217	0.0869
	5	0.0697	0.535	0.0697	0.139	0.223	0.268	0.214	0.0857
1	1	0.5	0.5	0.03125	0.15625	0.3125	0.3125	0.15625	0.03125
	2	0.390	0.390	0.107	0.267	0.312	0.212	0.0847	0.0169
	3	0.333	0.333	0.182	0.303	0.273	0.164	0.0655	0.0131
	4	0.302	0.302	0.242	0.302	0.242	0.145	0.058	0.0116
	5	0.285	0.285	0.285	0.285	0.228	0.137	0.055	0.011
0.5	1	0.667	0.333	0.1317	0.3292	0.3292	0.1646	0.0412	0.00412
	2	0.612	0.224	0.305	0.381	0.222	0.0754	0.0151	0.0015
	3	0.586	0.172	0.426	0.355	0.160	0.48	0.0096	0.00096
	4	0.572	0.144	0.509	0.318	0.127	0.0381	0.0076	0.00076
	5	0.564	0.128	0.564	0.282	0.113	0.0338	0.0068	0.0007

Table 3: Performance as a function of overlap

and the values of $\eta_j(S)$ are the rates at which server $j \in S$ gets activated by arrivals that are assigned to him, which can be recusively calculated by:

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

Let $\mathcal{P}(\{j_1, \dots, j_m\})$ be the set of all the permutations of $j_1, \dots, j_m$. Relevant to our derivation here is the property that for every $(\tilde{j}_1, \dots, \tilde{j}_m) \in \mathcal{P}(\{j_1, \dots, j_m\})$,

$$\eta_{\tilde{j}_1}(\{\tilde{j}_1\})\eta_{\tilde{j}_2}(\{\tilde{j}_1, \tilde{j}_2\}) \cdots \eta_{\tilde{j}_m}(\{\tilde{j}_1, \dots, \tilde{j}_m\}) = \eta_{j_1}(\{j_1\})\eta_{j_2}(\{j_1, j_2\}) \cdots \eta_{j_m}(\{j_1, \dots, j_m\}).$$

To distinguish between the two models we will now denote the stationary probabilities for the ALIS system by π_X.

Theorem 2 *The stationary distributions of $X(t)$ and of $Y(t)$ coincide in the sense that:*

$$\pi_Y(\{j_1,\dots,j_m\}) = \sum_{(\tilde{j}_1,\dots,\tilde{j}_m)\in\mathcal{P}(\{j_1,\dots,j_m\})} \pi_X(\tilde{j}_1,\dots,\tilde{j}_m). \tag{7}$$

Proof. We wish to show that for every subset $S=\{j_1,\dots,j_m\}$,

$$\begin{aligned}
&\pi_Y(\emptyset)\,\frac{\mu_{j_1}}{\eta_{j_1}(\{j_1\})}\,\frac{\mu_{j_2}}{\eta_{j_2}(\{j_1,j_2\})}\,\frac{\mu_{j_3}}{\eta_{j_3}(\{j_1,j_2,j_3\})}\cdots\frac{\mu_{j_m}}{\eta_{j_m}(S)}\\
&= \sum_{(\tilde{j}_1,\dots,\tilde{j}_m)\in\mathcal{P}(\{j_1,\dots,j_m\})} \pi_X(\emptyset)\,\frac{\mu_{\tilde{j}_1}}{\lambda_{\mathcal{C}(\tilde{j}_1)}}\,\frac{\mu_{\tilde{j}_2}}{\lambda_{\mathcal{C}(\{\tilde{j}_1,\tilde{j}_2\})}}\cdots\frac{\mu_{\tilde{j}_m}}{\lambda_{\mathcal{C}(\{\tilde{j}_1,\dots,\tilde{j}_m\})}}.
\end{aligned}$$

We note first that $\lambda_{C(S)}=\eta(S)$, and hence we need to verify that

$$\begin{aligned}
&\sum_{(\tilde{j}_1,\dots,\tilde{j}_m)\in\mathcal{P}(\{j_1,\dots,j_m\})} \Big(\eta(\{\tilde{j}_1\})\eta(\{\tilde{j}_1,\tilde{j}_2\})\cdots\eta(\{\tilde{j}_1,\dots,\tilde{j}_m\})\Big)^{-1} \\
&\qquad = \Big(\eta_{j_1}(\{j_1\})\eta_{j_2}(\{j_1,j_2\})\cdots\eta_{j_m}(\{j_1,\dots,j_m\})\Big)^{-1}.
\end{aligned}\tag{8}$$

Note that if (8) is valid for every subset $S=\{j_1,\dots,j_m\}$, then this implies equality of $\pi_Y(\emptyset)$ and $\pi_X(\emptyset)$. We prove (8) by induction on m. The case $m=1$ is immediate. Assume now that (8) holds for $m-1$. Then

$$\begin{aligned}
&\sum_{(\tilde{j}_1,\dots,\tilde{j}_m)\in\mathcal{P}(\{j_1,\dots,j_m\})} \Big(\eta(\{\tilde{j}_1\})\eta(\{\tilde{j}_1,\tilde{j}_2\})\cdots\eta(\{\tilde{j}_1,\dots,\tilde{j}_m\})\Big)^{-1}\\
&= \big(\eta(\{j_1,\dots,j_m\})\big)^{-1}\sum_{k=1}^{m}\ \sum_{(\tilde{j}_1,\dots,\tilde{j}_{m-1})\in\mathcal{P}(\{j_1,\dots,j_m\}\backslash\{j_k\})} \Big(\eta(\{\tilde{j}_1\})\eta(\{\tilde{j}_1,\tilde{j}_2\})\cdots\eta(\{\tilde{j}_1,\dots,\tilde{j}_{m-1}\})\Big)^{-1}\\
&= \big(\eta(\{j_1,\dots,j_m\})\big)^{-1}\sum_{k=1}^{m} \Big(\eta_{\tilde{j}_1}(\{\tilde{j}_1\})\eta_{\tilde{j}_2}(\{\tilde{j}_1,\tilde{j}_2\})\cdots\eta_{\tilde{j}_{m-1}}(\{\tilde{j}_1,\dots,\tilde{j}_{m-1}\})\Big)^{-1}\\
&= \big(\eta(\{j_1,\dots,j_m\})\big)^{-1}\sum_{k=1}^{m} \eta_{j_k}(\{j_1,\dots,j_m\})\\
&\qquad\qquad \Big(\eta_{\tilde{j}_1}(\{\tilde{j}_1\})\eta_{\tilde{j}_2}(\{\tilde{j}_1,\tilde{j}_2\})\cdots\eta_{\tilde{j}_{m-1}}(\tilde{j}_1,\dots,\tilde{j}_{m-1})\eta_{j_k}(\{j_k,\tilde{j}_1,\dots,\tilde{j}_{m-1}\})\Big)^{-1}\\
&= \big(\eta(\{j_1,\dots,j_m\})\big)^{-1}\sum_{k=1}^{m} \eta_{j_k}(\{j_1,\dots,j_m\})\Big(\eta_{j_1}(\{j_1\})\eta_{j_2}(\{j_1,j_2\})\cdots\eta_{j_m}(\{j_1,\dots,j_m\})\Big)^{-1}\\
&= \Big(\eta_{j_1}(\{j_1\})\eta_{j_2}(\{j_1,j_2\})\cdots\eta_{j_m}(\{j_1,\dots,j_m\})\Big)^{-1}.
\end{aligned}$$

The first equality is obtained by writing all the permutations of $\{j_1,\dots,j_m\}$ in terms of the permutations of $\{j_1,\dots,j_m\}\backslash\{j_k\}$ and adding them over k, and taking out of the sum the common term of $\big(\eta(\{j_1,\dots,j_m\})\big)^{-1}$. The second equality uses the induction hypothesis for $m-1$. In the third equality we multiply and divide each term in the sum by $\eta_{j_k}(\{j_1,\dots,j_m\})$. The fourth equality uses the property that $\eta_{\tilde{j}_1}(\{\tilde{j}_1\})\eta_{\tilde{j}_2}(\{\tilde{j}_1,\tilde{j}_2\})\cdots\eta_{\tilde{j}_m}(\{\tilde{j}_1,\dots,\tilde{j}_m\})$ has the same value for all permutations of $\{j_1,\ldots,j_m\}$. Finally, for the last equality we use $\eta(\{j_1,\ldots,j_m\}) = \sum_{k=1}^m \eta_{j_k}(\{j_1,\ldots,j_m\})$. This completes the proof. $\square$

Note: The process $X(t)$ is not time reversible, since new idle servers always join the current state as the last component, but servers which are activated may be from any component in the state, and are in fact the first server in the state which is compatible with an arriving customer, and may come from any position, mostly not from the last position.

6 Insensitivity

It is shown in [2] that the reversible loss system in insensitive. One can ask if the same property also holds for the loss system under ALIS regime. In this section we show that indeed our loss system under ALIS regime is insensitive, in that the stationary distribution depends on the service time distributions only through their means. This property can be expected because our system obeys partial balance. In fact we also show that for arbitrary service time distributions with finite means, at stationarity all busy servers have attained service and remaining service which are distributed according to the equilibrium distribution of the processing time. To show insensitivity we proceed as in [2]. We use the supplementary variable method, and our proof closely follows the 1957 proof of Sevastyanov [5], for the Erlang loss system.

We now assume that service times of server j are i.i.d. with distribution F_j with $F_j(0) = 0$, and finite mean $1/\mu_j$. We supplement the description of the state of the system at time t by specifying the attained service times of the busy servers. We let $Z(t)$ be the supplemented process, with $Z_j(t) = (X(t), W(t))$ where $X(t)$ is the list of idle servers at time t, ordered according to longest idle first, and $W(t)$ gives the attained service times of the busy servers, so that $w_j = *$ if $j \in X(t)$, to denote the server is idle, while $w_j \geq 0$ if server j is serving a customer, and the attained service time of the customer is w_j. Here we slightly abuse notation by writing $j \in X(t)$ where we mean that $j \in$ the set of elements of the list $X(t)$. We will denote by $P_t(s, w)$ the distribution of $Z(t)$,

$$P_t(s, w) = P(X(t) = s,\ W_j(t) \leq w_j,\ j \notin s)$$

For better readability we do not write that also $W_j(t) = *$, $j \in s$, which is understood from s. We will denote by $p_t(s, w)$ its density (which is shown in the proof to exist). We will denote by $P(s, w)$ and $p(s, w)$ the stationary distribution and density, respectively.

Proposition 3 *The process $Z(t)$ is ergodic with stationary probability density given by:*

$$p(s, w) = \pi(s) \prod_{j \notin s} \mu_j (1 - F_j(w_j)) \tag{9}$$

with $\pi(s)$ given in (1).

Proof. Let $P_t(s, w)$ be the distribution of $Z(t)$, with initial distribution P_0. It follows exactly as in Theorem 2 of [5] that for arbitrary P_0 and for any state (s, w), P_t has a density at the coordinates $w_j, j \notin s$, if $t > \max\{w_j : j \notin s\}$.

The process $Z(t)$ is a Markov process with transitions for large t and small Δ given by:

$$\begin{aligned}
&p_{t+\Delta}\Big((j_1,\ldots,j_m), w_j, j \notin \{j_1,\ldots,j_m\}\Big) = \\
&\quad p_t\Big((j_1,\ldots,j_m), w_j - \Delta, j \notin \{j_1,\ldots,j_m\}\Big)(1 - \lambda_{C(\{j_1,\ldots,j_m\})}\Delta) \prod_{j \notin S} \frac{1 - F_j(w_j)}{1 - F_j(w_j - \Delta)} \\
&\quad + \int_0^\infty p_t\Big((j_1,\ldots,j_{m-1}), W_{j_m} = y, w_j - \Delta, j \notin \{j_1,\ldots,j_m\}\Big)
\end{aligned}$$

$$\times \prod_{j \notin \{j_1,\dots,j_m\}} \frac{1-F_j(w_j)}{1-F_j(w_j-\Delta)} \frac{F_{j_m}(y+\Delta)-F_{j_m}(y)}{1-F_{j_m}(y)} dy + o(\Delta)$$

and for any $k \notin \{j_1,\dots,j_m\}$,

$$\begin{aligned}
&p_{t+\Delta}\Big((j_1,\dots,j_m), w_k = 0, w_j, j \notin \{k, j_1,\dots,j_m\}\Big)\Delta = \\
&\prod_{j \notin \{k,j_1,\dots,j_m\}} \frac{1-F_j(w_j)}{1-F_j(w_j-\Delta)} \Big[p_t\Big((k,j_1,\dots,j_m), w_j-\Delta, j \notin \{k,j_1,\dots,j_m\}\Big)\lambda_{C(k)}\Delta \\
&+p_t\Big((j_1,k,\dots,j_m), w_j-\Delta, j \notin \{k,j_1,\dots,j_m\}\Big)\lambda_{C(k)\setminus C(j_1)}\Delta + \cdots \\
&+p_t\Big((j_1,\dots,j_m,k), w_j-\Delta, j \notin \{k,j_1,\dots,j_m\}\Big)\lambda_{C(k)\setminus C(\{j_1,\dots,j_m\})}\Delta\Big] + o(\Delta).
\end{aligned}$$

Define now

$$p_t^*(s,w) = p_t(s,w) / \prod_{j \notin s} (1-F_j(w_j)),$$

to obtain:

$$\begin{aligned}
&p^*_{t+\Delta}\Big((j_1,\dots,j_m), w_j, j \notin \{j_1,\dots,j_m\}\Big) = \\
&\quad p^*_t\Big((j_1,\dots,j_m), w_j-\Delta, j \notin \{j_1,\dots,j_m\}\Big)(1-\lambda_{C(\{j_1,\dots,j_m\})}\Delta) \\
&\quad + \int_0^\infty p^*_t\Big((j_1,\dots,j_{m-1}), W_{j_m} = y, w_j-\Delta, j \notin \{j_1,\dots,j_m\}\Big) \\
&\qquad \times \big(F_{j_m}(y+\Delta)-F_{j_m}(y)\big)dy + o(\Delta)
\end{aligned}$$

and for any $k \notin \{j_1,\dots,j_m\}$,

$$\begin{aligned}
&p^*_{t+\Delta}\Big((j_1,\dots,j_m), w_k = 0, w_j, j \notin \{k,j_1,\dots,j_m\}\Big)\Delta = \\
&\quad p^*_t\Big((k,j_1,\dots,j_m), w_j-\Delta, j \notin \{k,j_1,\dots,j_m\}\Big)\lambda_{C(k)}\Delta \\
&\quad +p^*_t\Big((j_1,k,\dots,j_m), w_j-\Delta, j \notin \{k,j_1,\dots,j_m\}\Big)\lambda_{C(k)\setminus C(j_1)}\Delta + \cdots \\
&\quad +p^*_t\Big((j_1,\dots,j_m,k), w_j-\Delta, j \notin \{k,j_1,\dots,j_m\}\Big)\lambda_{C(k)\setminus C(\{j_1,\dots,j_m\})}\Delta + o(\Delta).
\end{aligned}$$

From these equations (and assuming that $p_t^*(z)$ is differentiable) we get a set of integro-differential equations:

$$\begin{aligned}
\frac{\partial p_t^*(s,w)}{\partial t} + \sum_{j \notin \{j_1,\dots,j_m\}} \frac{\partial p_t^*(s,w)}{\partial w_j} &= -\lambda_{C(\{j_1,\dots,j_m\})} p_t^*(s,w) \\
&\quad + \int_0^\infty p_t^*\big((j_1,\dots,j_{m-1}), W_{j_m} = y, w\big) dF_{j_m}(y)
\end{aligned}$$

with boundary conditions for all $k \notin \{j_1,\dots,j_m\}$:

$$\begin{aligned}
&p_t^*\big((j_1,\dots,j_m), w_k = 0, w\big) = p_t^*\big((k,j_1,\dots,j_m), w\big)\lambda_{C(k)} \\
&\quad + p_t^*\big((j_1,k,\dots,j_m), w\big)\lambda_{C(k)\setminus C(j_1)} + \cdots + p_t^*\big((j_1,\dots,j_m,k), w\big)\lambda_{C(k)\setminus C(\{j_1,\dots,j_m\})}.
\end{aligned}$$

In stationarity the derivatives with respect to t cancel, so that we have:

$$\sum_{j\notin\{j_1,\dots,j_m\}} \frac{\partial p^*(s,w)}{\partial w_j} = -\lambda_{C(\{j_1,\dots,j_m\})} p^*(s,w)$$
$$+\int_0^\infty p^*\big((j_1,\dots,j_{m-1}),W_{j_m}=y,w\big)dF_{j_m}(y)$$

with boundary conditions:

$$p^*\big((j_1,\dots,j_m),w_k=0,w\big) = p^*\big((k,j_1,\dots,j_m),w\big)\lambda_{C(k)}$$
$$+p^*\big((j_1,k,\dots,j_m),w\big)\lambda_{C(k)\setminus C(j_1)} + \cdots + p^*\big((j_1,\dots,j_m,k),w\big)\lambda_{C(k)\setminus C(\{j_1,\dots,j_m\})}.$$

We now put in the trial solution

$$p^*(s,w) = \pi(s)\prod_{j\notin s}\mu_j.$$

Note that w_j do not appear in this trial solution, so the partial differentials in the first equation are zero, and the integral is over a constant. We obtain for the first equation:

$$\pi(j_1,\dots,j_m) \prod_{j\notin\{j_1,\dots,j_m\}} \mu_j \;\; \lambda_{C(\{j_1,\dots,j_m\})} = \pi(j_1,\dots,j_{m-1})\; \mu_{j_m} \prod_{j\notin\{j_1,\dots,j_m\}} \mu_j,$$

which is exactly the partial balance equation (3), satisfied by π for any s. In the second equation we obtain, for any $(j_1,\dots,j_m)$ and $k\notin\{j_1,\dots,j_m\}$:

$$\pi(j_1,\dots,j_m)\mu_k \prod_{j\notin\{k,j_1,\dots,j_m\}} \mu_j$$
$$= \big[\pi(k,j_1,\dots,j_m)\lambda_{C(k)} + \pi(j_1,k,\dots,j_m)\lambda_{C(k)\setminus C(j_1)} + \cdots$$
$$+\pi(j_1,\dots,j_m,k)\lambda_{C(k)\setminus C(\{j_1,\dots,j_m\})}\big] \prod_{j\notin\{k,j_1,\dots,j_m\}} \mu_j.$$

which is exactly the partial balance equation (4) satisfied by π for each s and $k\notin s$.

This confirms that (9) is a stationary density for the Markov process $Z(t)$. It can now be shown exactly as in [5] that $Z(t)$ is ergodic with a unique stationary density. □

References

[1] Adan, I.J.B.F., Weiss, G. (2011) Exact FCFS matching rates for two infinite multi-type sequences. *Operations Research* to appear.

[2] Adan, I.J.B.F., Hurkens, C., Weiss, G. (2010) A reversible loss system with multitype customers and multitype servers. *Probability in Engineering and Informational Sciences* 24:535–548.

[3] Busic, A., Gupta, V. and Mairesse, J. (2010) Stability of the bipartite matching model, Preprint, arxiv:1003.3477v1 [cs.DM].

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

[5] Sevastyanov, B.A. (1957) An ergodic theorem for Markov processes and its application to telephone systems with refusals *Theory of Probability and its Applications* 2:104–112.

[6] Talreja, R. and Whitt, W. (2008) Fluid Models for Overloaded Multi-class many-service queueing systems with FCFS routing. *Management Science* 54:1513–1527.

[7] Visschers, J.W.C.H., Adan, I.J.B.F., Weiss, G. (2011). A product form solution to a system with multi-type customers and multi-type servers. *Queueing Systems* to appear.