




Centrum voor Wiskunde en Informatica

View metadata, citation and similar papers at core.ac.uk

brought to you by  CORE

provided by CWI's Institutions

REPORTRAPPORT

PNA

Probability, Networks and Algorithms



Probability, Networks and Algorithms

Stochastic bounds for two-layer loss systems

M. Jonckheere, L.S. Leskelä

REPORT PNA-R0708 AUGUST 2007

Centrum voor Wiskunde en Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO). CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

Copyright © 2007, Stichting Centrum voor Wiskunde en Informatica
P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

Stochastic bounds for two-layer loss systems

ABSTRACT

This paper studies multiclass loss systems with two layers of servers, where each server at the first layer is dedicated to a certain customer class, while the servers at the second layer can handle all customer classes. The routing of customers follows an overflow scheme, where arriving customers are preferentially directed to the first layer. Stochastic comparison and coupling techniques are developed for studying how the system is affected by packing of customers, altered service rates, and altered server configurations. This analysis leads to easily computable upper and lower bounds for the performance of the system.

2000 Mathematics Subject Classification: 60K25; 60E15; 68M20; 90B15; 90B22

Keywords and Phrases: multiclass loss system; overflow routing; maximum packing; stochastic order; preorder; coupling

Stochastic bounds for two-layer loss systems

M. Jonckheere and L. Leskelä

July 13, 2007

Abstract

This paper studies multiclass loss systems with two layers of servers, where each server at the first layer is dedicated to a certain customer class, while the servers at the second layer can handle all customer classes. The routing of customers follows an overflow scheme, where arriving customers are preferentially directed to the first layer. Stochastic comparison and coupling techniques are developed for studying how the system is affected by packing of customers, altered service rates, and altered server configurations. This analysis leads to easily computable upper and lower bounds for the performance of the system.

Keywords: multiclass loss system, overflow routing, maximum packing, stochastic order, preorder, coupling

AMS Subject Classification: 60K25, 60E15, 68M20, 90B15, 90B22

1 Introduction

This paper studies multiclass loss systems with two layers of servers, where each server at the first layer is dedicated to a certain customer class, and the servers at the second layer can handle all customer classes. Arriving customers are routed to vacant servers in one of the layers, with preference given to the first layer; or rejected otherwise. This policy is commonly referred to as overflow routing.

Layered networks with overflow routing are commonly used in telecommunications services, because different layers of service may increase the system capacity. In *wireless communication networks* for instance, the servers at the first layer correspond to radio channels dedicated to a small geographical area (microcell), and the second layer represents available radio channels in a larger area covering several microcells; in *telephone call centers*, the first

layer consists of call agents trained to handling certain types of phone calls, and the second layer represents call agents who are cross-trained to deal with all types of calls.

The analysis of multilayer loss systems is challenging even under the simplest statistical assumptions, because the distributions of the overflow processes from the first layer are complex, and the direct numerical computation of the stationary distribution is unfeasible even for relatively small systems (Louth, Mitzenmacher, and Kelly [8]). Hence, approximative methods are needed for performance analysis (see Kelly [7] for a broad overview). Classical approximation techniques such as the equivalent random method and the Hayward–Fredericks method [14], and the recently introduced hyperexponential decomposition (Franx, Koole, and Pot [3]), are based on parametrically modeling the overflow processes from the first layer by simpler processes. These methods have been observed to produce good approximations for many choices of system parameters. However, they may require considerable amounts of computation, and it is not clear whether they remain accurate over the full parameter range.

The goal of this paper is to approximate the system via upper and lower bounds that are easy to compute numerically, and conservative in the sense that the true performance remains between the bounds for all choices of system parameters. To construct the upper bound, we modify the system by redirecting customers from the second layer into the first layer as soon as servers become vacant. This so-called maximum packing policy causes the number of customers per class to have a product-form stationary distribution (Everitt and Macfadyen [2]). The lower bound is constructed by moving all servers from the second layer into the first, this way reducing the system into a product of independent Erlang loss models.

The main tools for proving the validity of the bounds are (i) Massey’s theorem [9] characterizing the comparability of two Markov jump processes; and (ii) stochastic coupling, where versions of the queue-length processes for the original and for the reference model are constructed in such a way that the difference of the two processes remains positive with probability one. In the context of loss systems, coupling-based stochastic bounds have successfully been derived among others by Whitt [13], who analyzed several single-class queueing systems; Nain [11], who focused on multiclass single-layer loss systems; and Hordijk and Ridder [4], who studied a special case of the two-layer loss system where the first layer is fully dedicated to a single customer class. This paper extends some of the above results to general multiclass two-layer loss systems, the main contribution being in showing that maximum packing leads to upper bounds for the time-dependent and

stationary distributions of the number of customers in the system. In the special case where the first layer is fully dedicated to a single customer class, this result improves the upper bound obtained by Hordijk and Ridder [4].

The paper is organized as follows. Section 2 introduces the model details and notation. In Section 3 we prove a preliminary comparison result that is key to analyzing the monotonicity of the system. Section 4 analyzes how the time-dependent distribution of the system is affected by maximum packing, different server configurations, and altered service rates, and in Section 5 we carry out a similar analysis for the system in steady state. Section 6 concludes the paper.

2 Model description

2.1 Two-layer loss system with overflow routing

We consider a loss system with K customer classes and two layers of servers, where layer 1 contains M_k servers dedicated to class k , and layer 2 consists of N servers capable of serving all customer classes. Arriving class- k customers are routed to vacant servers in one of the layers, with preference given to layer 1; or rejected otherwise (Figure 1). For analytical tractability, we assume that the interarrival times and the service requirements of class- k customers are exponentially distributed with parameters λ_k and μ_k , respectively, and that all these random variables across all customer classes are independent.

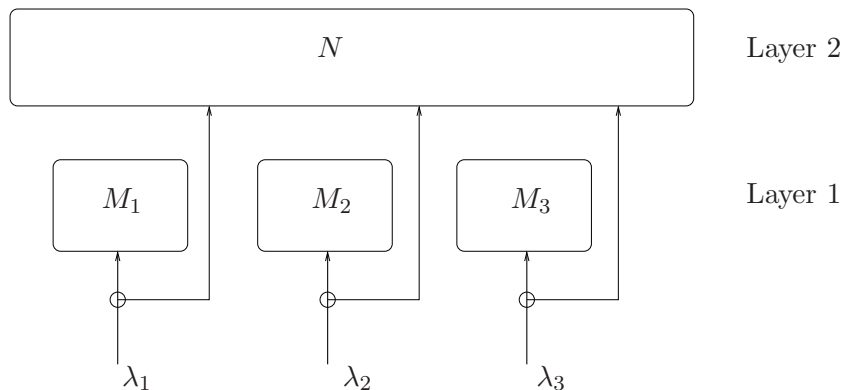


Figure 1: Two-layer loss network with three customer classes.

Denote by $X_{i,k}$ the number of class- k customers being served at layer i .

The system is described by the random vector $X = (X_{i,k})$ taking values in

$$S = \{x \in \mathbb{Z}_+^K \times \mathbb{Z}_+^K : x_{1,k} \leq M_k \forall k, |x|_2 \leq N\}, \quad (1)$$

where $|x|_2 = \sum_{k=1}^K x_{2,k}$. Let us also denote by $e_{i,k}$ the unit vector in $\mathbb{Z}_+^K \times \mathbb{Z}_+^K$ corresponding to the coordinate direction (i, k) . Moreover, define the sets

$$A_{1,k} = \{x \in S : x_{1,k} < M_k\}, \quad (2)$$

$$A_{2,k} = \{x \in S : x_{1,k} = M_k, |x|_2 < N\}, \quad (3)$$

$$B_k = \{x \in S : x_{1,k} = M_k, |x|_2 = N\}. \quad (4)$$

The set $A_{i,k}$ represents the set of states where an arriving class- k customer is assigned to a layer- i server, and B_k is the set of states where arriving class- k customers are rejected. The process X is a continuous-time Markov process on S with the upward transitions $x \mapsto x + e_{i,k}$ at rate $\lambda_{i,k}(x)$, and downward transitions $x \mapsto x - e_{i,k}$ at rate $\phi_{i,k}(x)$, where

$$\begin{aligned} \lambda_{i,k}(x) &= \lambda_k 1(x \in A_{i,k}), \\ \phi_{i,k}(x) &= \mu_k x_{i,k}. \end{aligned} \quad (5)$$

2.2 Maximum packing

To approximate the original two-layer loss system, we consider a modification of the system, where customers are redirected from layer 2 to layer 1 as soon as servers become vacant. This corresponds to the so-called maximum packing policy introduced by Everitt and Macfadyen [2]. The queue-length process X^{mp} for this system is a continuous-time Markov process on S with the upward transitions $x \mapsto x + e_{i,k}$ at rate $\lambda'_{i,k}(x)$, and downward transitions $x \mapsto x - e_{i,k}$ at rate $\phi'_{i,k}(x)$, where

$$\begin{aligned} \lambda'_{i,k}(x) &= \lambda_k 1(x \in A_{i,k}), \quad i = 1, 2, \\ \phi'_{1,k}(x) &= \mu_k x_{1,k} 1(x_{2,k} = 0), \\ \phi'_{2,k}(x) &= \mu_k x_{1,k} 1(x_{2,k} > 0) + \mu_k x_{2,k}. \end{aligned} \quad (6)$$

Remark 1. A remarkable property of the maximum packing policy is that all states outside the set $S^{\text{mp}} = \bigcap_{k=1}^K \{x \in S : x_{1,k} = M_k \text{ or } x_{2,k} = 0\}$ are transient for X^{mp} . Moreover, note that for $x \in S^{\text{mp}}$, $x_{1,k} = M_k$ if and only if $x_{1,k} + x_{2,k} \geq M_k$, which implies that

$$\begin{aligned} x_{1,k} &= (x_{1,k} + x_{2,k}) \wedge M_k, \\ x_{2,k} &= (x_{1,k} + x_{2,k} - M_k)^+. \end{aligned} \quad (7)$$

As a consequence, the aggregate queue-length process $(X_{1,k}^{\text{mp}} + X_{2,k}^{\text{mp}})_{k=1}^K$, if started in S^{mp} , is equal in distribution to \hat{X}^{mp} , where $\hat{X}^{\text{mp}} = (\hat{X}_1^{\text{mp}}, \dots, \hat{X}_K^{\text{mp}})$ is a Markov process on $\hat{S}^{\text{mp}} = \{\hat{x} \in \mathbb{Z}_+^K : \sum_k (\hat{x}_k - M_k)^+ \leq N\}$ generated by the transitions

$$\hat{x} \mapsto \begin{cases} \hat{x} + e_k, & \text{at rate } \lambda_k 1(\hat{x} + e_k \in \hat{S}^{\text{mp}}), \\ \hat{x} - e_k, & \text{at rate } \phi_k(\hat{x}) = \mu_k \hat{x}_k. \end{cases}$$

The structure of the above transition rates implies that the stationary distribution of \hat{X}^{mp} is a product of Poisson distributions truncated to \hat{S}^{mp} , which is easy to compute numerically. The stationary distribution of X^{mp} can then be recovered from that of \hat{X}^{mp} using the equalities (7).

3 Preliminary result

This section establishes a general result that allows us to compare two processes taking values in $S \subset \mathbb{Z}_+^K \times \mathbb{Z}_+^K$ with respect to a specific preorder. This preorder, tailored to fit the transition rates of the type in (5), is defined by $x \prec y$, if $x_{1,k} \leq y_{1,k}$ for all k and $|x| \leq |y|$, where $|x| = \sum_{i,k} x_{i,k}$. Recall that the usual stochastic order [10] between random variables is defined by $X \leq_{\text{st}} Y$ if $\mathbb{E} f(X) \leq \mathbb{E} f(Y)$ for all nonnegative increasing functions f .

Consider a continuous-time Markov process X on $S \subset \mathbb{Z}_+^K \times \mathbb{Z}_+^K$ generated by the transitions

$$x \mapsto \begin{cases} x + e_{i,k} & \text{at rate } \lambda_{i,k}(x), \\ x - e_{i,k} & \text{at rate } \phi_{i,k}(x), \end{cases}$$

$i \in \{1, 2\}$, $k \in \{1, \dots, K\}$, where $\lambda_{i,k}$ and $\phi_{i,k}$ are bounded nonnegative functions on S . For consistency, we assume here that $\lambda_{i,k}(x) = 0$ for all $x \in S$ such that $x + e_{i,k} \notin S$ and $\phi_{i,k}(x) = 0$ for all $x \in S$ such that $x - e_{i,k} \notin S$. We assume that Y is a similar process with state-dependent transition rates $\lambda'_{i,k}$ and $\phi'_{i,k}$.

Theorem 1. *Let X and Y be continuous-time Markov processes on S having upward transition rates $\lambda_{i,k}$ and $\lambda'_{i,k}$, and downward transition rates $\phi_{i,k}$ and $\phi'_{i,k}$, respectively. Assume that:*

(i) *For all $x, y \in S$ such that $x \prec y$ and $x_{1,k} = y_{1,k}$,*

$$\lambda_{1,k}(x) \leq \lambda'_{1,k}(y), \tag{8}$$

$$\phi_{1,k}(x) \geq \phi'_{1,k}(y). \tag{9}$$

(ii) For all $x, y \in S$ such that $x \prec y$ and $|x| = |y|$,

$$\sum_{i,k} \lambda_{i,k}(x) \leq \sum_{i,k} \lambda'_{i,k}(y), \quad (10)$$

$$\sum_{i,k} \phi_{i,k}(x) \geq \sum_{i,k} \phi'_{i,k}(y). \quad (11)$$

Assuming that the initial states satisfy $X(0) \prec Y(0)$, it then follows that $X_{1,k}(t) \leq_{\text{st}} Y_{1,k}(t)$ for all k and t , and $|X(t)| \leq_{\text{st}} |Y(t)|$ for all t .

Proof. Denote the infinitesimal generators of X and Y by p and q , respectively. Recall that $U \subset S$ is called an *upper set*, if $x \in U$ and $x \prec y$ implies $y \in U$, and $V \subset S$ is called a *lower set*, if the complement V^c of V is an upper set. Using a result of Massey [9, Theorem 5.3]¹, it suffices to verify that $p(x, U) \leq q(y, U)$ for all $x \prec y$ and for all upper sets U such that either $x \in U$ or $y \notin U$. Because $p(x, U) = -p(x, U^c)$ for all $x \in U$, this condition is equivalent to showing that for all $x \prec y$,

$$\sum_{i,k} \lambda_{i,k}(x) 1(x + e_{i,k} \in U) \leq \sum_{i,k} \lambda'_{i,k}(y) 1(y + e_{i,k} \in U) \quad (12)$$

for all upper sets U such that $x \notin U, y \notin U$, and

$$\sum_{i,k} \phi_{i,k}(x) 1(x + e_{i,k} \in V) \geq \sum_{i,k} \phi'_{i,k}(y) 1(y + e_{i,k} \in V) \quad (13)$$

for all lower sets V such that $x \notin V, y \notin V$.

Assume $x \prec y$ and choose an upper set U such that $x \notin U, y \notin U$. To verify the validity of (12), let us consider separately the cases $|x| < |y|$ and $|x| = |y|$. Assume first $|x| < |y|$. Then $x + e_{1,k} \prec y$ for all k such that $x_{1,k} < y_{1,k}$, and $x + e_{2,k} \prec y$ for all k . Hence because U is an upper set and $y \notin U$, it follows that $x + e_{1,k} \in U$ only if $x_{1,k} = y_{1,k}$, and $x + e_{2,k} \notin U$ for all k . Thus,

$$\sum_{i,k} \lambda_{i,k}(x) 1(x + e_{i,k} \in U) = \sum_{k: x_{1,k} = y_{1,k}} \lambda_{1,k}(x) 1(x + e_{1,k} \in U). \quad (14)$$

Moreover, using inequality (8), and noting that $x + e_{1,k} \prec y + e_{1,k}$ for all k such that $y + e_{1,k} \in S$, we see that for all k such that $x_{1,k} = y_{1,k}$,

$$\lambda_{1,k}(x) 1(x + e_{1,k} \in U) \leq \lambda'_{1,k}(y) 1(y + e_{1,k} \in U). \quad (15)$$

¹Massey formulated his result for partially ordered spaces, but all the proofs in his paper [9] remain valid also for preorders that are not antisymmetric.

Substituting (15) into (14) shows the validity of (12).

Let us next focus on the case $|x| = |y|$. Note first that if $x + e_{1,l} \in U$ for some l such that $x_{1,l} < y_{1,l}$, or $x + e_{2,l} \in U$ for some l , then $y + e_{i,k} \in U$ for all i and k . Hence it follows that the right-hand side of (12) equals $\sum_{i,k} \lambda'_{i,k}(y)$, which in light of assumption (10) guarantees the validity of (12). On the other hand, if $x + e_{2,k} \notin U$ for all k , and $x_{1,k} = y_{1,k}$ for all k such that $x + e_{1,k} \in U$, then equation (14) holds. Assumption (8) again implies (15), which together with (14) shows the validity of (12).

The proof is completed by carrying out an analogous reasoning for lower sets, which shows that assumptions (9) and (11) imply (13). \square

4 Stochastic comparisons

This section contains the main results for analyzing the time-dependent distribution of the system. Assuming first that all service rates across different customer classes are equal, we study how the system is affected by maximum packing (Section 4.1) and different server configurations (Section 4.2). Section 4.3 provides a monotonicity result that allows to extend the analysis to the case where the service rates are not assumed equal, and Section 4.4 describes bounds for the per-class number of customers in the system.

4.1 Maximum packing

Let X be the queue-length process of the two-layer loss system defined in Section 2.1, and denote by X^{mp} the queue-length process corresponding to the maximum packing policy, as defined in Section 2.2. Recall from Section 3 that the preorder $x \prec y$ is defined by $x_{1,k} \leq y_{1,k}$ for all k and $|x| \leq |y|$.

Theorem 2. *Assume that all service rates μ_k are equal. Given that the initial states satisfy $X(0) \prec X^{\text{mp}}(0)$, it then follows that $X_{1,k}(t) \leq_{\text{st}} X_{1,k}^{\text{mp}}(t)$ for all k and t , and $|X(t)| \leq_{\text{st}} |X^{\text{mp}}(t)|$ for all t .*

Remark 2. Example 1 in Section 5.2 shows that the statement of Theorem 2 may not be true, if the service rates are not assumed equal.

The proof of Theorem 2 is based on the following lemma.

Lemma 1. *The transition rates $\lambda_{i,k}(x)$ defined in (5) satisfy:*

(i) *For all $x \prec y$ and for all k such that $x_{1,k} = y_{1,k}$,*

$$\lambda_{1,k}(x) \leq \lambda_{1,k}(y). \tag{16}$$

(ii) For all $x \prec y$ and for all k such that $|x| = |y|$,

$$\sum_{i,k} \lambda_{i,k}(x) \leq \sum_{i,k} \lambda_{i,k}(y). \quad (17)$$

Proof. The inequality (16) is clear, because $\lambda_{1,k}(x)$ only depends on $x_{1,k}$. Assume next that $x \prec y$ and $\sum_{i,k} x_{i,k} = \sum_{i,k} y_{i,k}$. Assume that $y \in B_k$ for some k , where B_k is defined in (4). Then $|y_2| = N$, which implies that $|x_2| = N$ and $x_{1,l} = y_{1,l}$ for all l . Thus $x \in B_k$. We may thus conclude that for all k , $1(x \notin B_k) \leq 1(y \notin B_k)$. Hence it follows that

$$\sum_{i,k} \lambda_{i,k}(x) = \sum_k \lambda_k 1(x \notin B_k) \leq \sum_k \lambda_k 1(y \notin B_k) = \sum_{i,k} \lambda_{i,k}(y),$$

which shows the validity of (17). \square

Proof of Theorem 2. Let $\lambda_{i,k}(x)$ and $\phi_{i,k}(x)$ be the transition rates of X as defined in (5), and let $\lambda'_{i,k}(x)$ and $\phi'_{i,k}(x)$ be the corresponding rates for X^{mp} as defined in (6). Because $\lambda'_{i,k}(x) = \lambda_{i,k}(x)$ for all x , the validity of (8) and (10) in Theorem 1 follow by Lemma 1. For the downward transitions, note that for all $x \prec y$ such that $x_{1,k} = y_{1,k}$ for some k , $\phi_{1,k}(x) = \mu_1 x_{1,k} = \mu_1 y_{1,k} \geq \mu_1 y_{1,k} 1(y_{2,k} = 0) = \phi'_{1,k}(y)$. Moreover, for all $x \prec y$ such that $|x| = |y|$,

$$\sum_k (\phi_{1,k}(x) + \phi_{2,k}(x)) = \mu_1 |x| = \mu_1 |y| = \sum_k (\phi'_{1,k}(y) + \phi'_{2,k}(y)),$$

so conditions (9) and (11) of Theorem 1 are valid. Hence Theorem 1 yields the claim. \square

4.2 Different server configurations

This section studies the effect of moving one server from layer 1 to layer 2. As in Section 2.1, we denote by X the queue-length process of the system with server configuration $M = (M_1, \dots, M_K)$ in layer 1, and N servers in layer 2. Let Y be the queue-length process of the modified system where one class- k server from layer 1 has been replaced by a server in layer 2. We assume $k = 1$ without loss of generality. Let $M' = (M_1 - 1, M_2, \dots, M_K)$ and $N' = N + 1$, and define the sets S' , $A'_{1,k}$ and B'_k as in (1)–(4) with M and N replaced by M' and N' , respectively. Then Y is a Markov process on S' having transition rates of the form (5) with $A_{i,k}$ replaced by $A'_{i,k}$.

Let us denote by $x_2 = \sum_k x_{2,k}$ the number of customers being served at layer 2. Assuming that all service rates μ_k are equal, it follows that the process $(X_{1,1}, \dots, X_{1,K}; X_2)$ is Markov. With a slight abuse of notation, we will redefine the state space by $S = \{(x_{1,1}, \dots, x_{1,K}; x_2) \in \mathbb{Z}_+^K \times \mathbb{Z}_+ : x_{1,k} \leq M_k \text{ for all } k, x_2 \leq N\}$, and denote by e_2 the unit vector in $\mathbb{Z}_+^K \times \mathbb{Z}_+$ corresponding to the last coordinate. We will redefine the sets $A_{i,k}, B_k, A'_{i,k}, B'_k$, and S' in a similar way, identifying $|x|_2$ with x_2 .

Theorem 3. *Assume that all service rates μ_k are equal, and let $\Delta = \{0, e_2, e_2 - e_{1,1}, 2e_2 - e_{1,1}\}$. Then, assuming that the initial states satisfy $Y(0) - X(0) \in \Delta$, it follows that $|X(t)| \leq_{\text{st}} |Y(t)|$ for all t .*

Proof. Because $|x| \leq |y|$ for all $x \in S$ and $y \in S'$ such that $y - x \in \Delta$, it is sufficient to construct a coupling [12] of X and Y that takes values in $S_\Delta = \{(x, y) \in S \times S' : y - x \in \Delta\}$. Let (\tilde{X}, \tilde{Y}) be a continuous-time Markov process on S_Δ generated by the joint arrivals

$$(x, y) \mapsto (x + e_{1,k}, y + e_{1,k}) \quad \text{at rate } \lambda_k 1(x \in A_{1,k}, y \in A'_{1,k}), \quad (18)$$

$$(x, y) \mapsto (x + e_{1,k}, y + e_2) \quad \text{at rate } \lambda_k 1(x \in A_{1,k}, y \in A'_{2,k}), \quad (19)$$

$$(x, y) \mapsto (x + e_{1,k}, y) \quad \text{at rate } \lambda_k 1(x \in A_{1,k}, y \in B'_k), \quad (20)$$

$$(x, y) \mapsto (x + e_2, y + e_2) \quad \text{at rate } \sum_l \lambda_l 1(x \in A_{2,l}, y \in A'_{2,l}), \quad (21)$$

$$(x, y) \mapsto (x + e_2, y) \quad \text{at rate } \sum_l \lambda_l 1(x \in A_{2,l}, y \in B'_l), \quad (22)$$

$$(x, y) \mapsto (x, y + e_2) \quad \text{at rate } \sum_l \lambda_l 1(x \in B_l, y \in A'_{2,l}), \quad (23)$$

and joint departures

$$(x, y) \mapsto (x - e_{1,k}, y - e_{1,k}) \quad \text{at rate } \mu_1 y_{1,k}, \quad (24)$$

$$(x, y) \mapsto (x - e_{1,1}, y - e_2) \quad \text{at rate } \mu_1 (x_{1,1} - y_{1,1}), \quad (25)$$

$$(x, y) \mapsto (x - e_2, y - e_2) \quad \text{at rate } \mu_1 x_2, \quad (26)$$

$$(x, y) \mapsto (x, y - e_2) \quad \text{at rate } \mu_1 (y_{1,1} + y_2 - x_{1,1} - x_2). \quad (27)$$

Observe that all transition rates above are nonnegative, because $y_{1,1} \leq x_{1,1}$ and $y_{1,1} + y_2 \geq x_{1,1} + x_2$, whenever $y - x \in \Delta$. To ensure that the transitions define a generator of a Markov process on S_Δ , we need to verify that $y' - x' \in \Delta$ for all transitions $(x, y) \mapsto (x', y')$, where $y - x \in \Delta$. This is obvious for transitions (18), (21), (24), and (26), because in these cases $y' - x' = y - x$. Let us consider the remaining cases one-by-one:

- If transition (19) occurs, then $k = 1$, because $y_{1,k} = x_{1,k}$ for all $k \neq 1$. Then $x_{1,1} < M_1$ and $y_{1,1} = M_1 - 1$, so it follows that either $y - x = 0$ or $y - x = e_2$. In both cases, $y' - x' \in \Delta$.

- If transition (20) occurs, then again $k = 1$. Then $x_{1,1} < M_1$ and $y_{1,1} = M_1 - 1$, which implies $y_{1,1} = x_{1,1}$. Moreover, $y_2 = N + 1$, which is only possible if $y_2 = x_2 + 1$. Hence $y - x = e_2$, so that $y' - x' = e_2 - e_{1,1} \in \Delta$.
- If transition (22) occurs, then $x \in A_{2,l}$ and $y \in B'_l$ for some l . Then $x_2 < N$ and $y_2 = N + 1$, which implies that $y - x = 2e_2 - e_{1,1}$. Hence $y' - x' = e_2 - e_{1,1} \in \Delta$.
- If transition (23) occurs, then $x \in B_l$ and $y \in A'_{2,l}$ for some l . Then $x_2 = N$ and $y_2 < N + 1$, so it follows that $y_2 = x_2$. Hence $y - x = 0$, and thus $y' - x' = e_2 \in \Delta$.
- If transition (25) occurs, then $y_{1,1} < x_{1,1}$. Because $y - x \in \Delta$, this implies that either $y - x = e_2 - e_{1,1}$, so that $y' - x' = 0$; or $y - x = 2e_2 - e_{1,1}$, so that $y' - x' = e_2$.
- If transition (27) occurs, then $y_{1,1} + y_2 - x_{1,1} - x_2 > 0$. Because $y - x \in \Delta$, it follows that either $y - x = e_2$, so that $y' - x' = 0$; or $y - x = 2e_2 - e_{1,1}$, so $y' - x' = e_2 - e_{1,1}$.

Hence, all transitions map S_Δ into S_Δ , and the process (\tilde{X}, \tilde{Y}) is well-defined.

To show that (\tilde{X}, \tilde{Y}) is a coupling of X and Y , we must verify that the marginal transition rates of (\tilde{X}, \tilde{Y}) match with the transition rates of X and Y . Note first that the sum of transition rates such that $x \mapsto x + e_{1,k}$ is equal to $\lambda_k 1(x \in A_{1,k})$. Next, observe that $x \in A_{2,l}$ and $y - x \in \Delta$ imply that $y \notin A'_{1,l}$. Hence the sum of transition rates where $x \mapsto x + e_2$ is equal to

$$\sum_l \lambda_l 1(x \in A_{2,l}, y \in A'_{2,l} \cup B'_l) = \sum_l \lambda_l 1(x \in A_{2,l}).$$

Further, because the sum of all transition rates such that $x \mapsto x - e_{1,k}$ equals $\mu_1 x_{1,k}$ for all k , and the corresponding sum for $x \mapsto x - e_2$ is equal to $\mu_1 x_2$, we may conclude that the transitions of \tilde{X} and X occur at the same rates.

Turning the attention to the rates of \tilde{Y} , note that $y - x \in \Delta$ and $y \in A'_{1,1}$ imply that $y_{1,1} < M_1 - 1$ and $x_{1,1} \leq y_{1,1} + 1$, so it follows that $x \in A_{1,1}$. Moreover, $y - x \in \Delta$ and $y \in A'_{1,k}$ for $k \neq 1$ imply that $x_{1,k} = y_{1,k} < M_k$, so $x \in A_{1,k}$. Hence the total rate of transitions where $y \mapsto y + e_{1,k}$ is equal to $\lambda_k 1(x \in A_{1,k}, y \in A'_{1,k}) = \lambda_k 1(y \in A'_{1,k})$. Further, because the net rate of transitions where $y \mapsto y + e_2$ is equal to $\sum_l \lambda_l 1(y \in A'_{2,l})$, and because the corresponding net rates for $y \mapsto y - e_{1,k}$ and $y \mapsto y - e_2$ are equal to $\mu_1 y_{1,k}$

and $\mu_1 y_2$, respectively, we conclude that the transitions of \tilde{Y} and Y occur at the same rates. Hence, the process (\tilde{X}, \tilde{Y}) is a coupling of X and Y . \square

4.3 Monotonicity with respect to service rates

The results in Sections 4.1 and 4.2 were proved under the assumption that all service rates are equal. The following theorem describes a monotonicity property that allows to compare systems not satisfying this assumption. Denote by X the queue-length process of the two-layer loss system defined in Section 2.1, and let X^- and X^+ be modifications of the system with all service rates set to $\mu_{\max} = \max \mu_k$ and $\mu_{\min} = \min \mu_k$, respectively. Recall that the preorder $x \prec y$ is defined by $x_{1,k} \leq y_{1,k}$ for all k and $|x| \leq |y|$.

Theorem 4. *Assume that the initial states satisfy $X^-(0) \prec X(0) \prec X^+(0)$. Then for all t ,*

$$X_{1,k}^-(t) \leq_{\text{st}} X_{1,k}(t) \leq_{\text{st}} X_{1,k}^+(t)$$

for all k , and

$$|X^-(t)| \leq_{\text{st}} |X(t)| \leq_{\text{st}} |X^+(t)|.$$

Remark 3. A simpler comparison statement, such as $|X(t)| \leq_{\text{st}} |X^+(t)|$ given $|X(0)| \leq |X^+(0)|$, is not true in general. Using Massey's [9] criteria for the preorder $|x| \leq |y|$, it is not hard to check that a necessary condition for the above property is that $\sum_{i,k} \lambda_{i,k}(x) = \sum_{i,k} \lambda_{i,k}(y)$ whenever $|x| = |y|$. This equality may fail for instance for $x = \sum_k M_k e_{1,k} + (N-1)e_{2,1}$ and $y = x - e_{1,1} + e_{2,1}$.

Proof of Theorem 4. Note that X^+ has the same upward transitions as X and downward transitions $\phi'_{1,k}(x) = \mu_{\min} x_{1,k}$, and $\phi'_{2,k}(x) = \mu_{\min} x_{2,k}$. Now for all $x \prec y$ such that $x_{1,k} = y_{1,k}$ for some k , $\mu_k x_{1,k} \geq \mu_{\min} x_{1,k} = \mu_{\min} y_{1,k}$, and for all $x \prec y$ such that $|x| = |y|$,

$$\sum_k \mu_k (x_{1,k} + x_{2,k}) \geq \mu_{\min} \sum_k (x_{1,k} + x_{2,k}) = \mu_{\min} \sum_{i,k} (y_{1,k} + y_{2,k}),$$

so conditions (9) and (11) of Theorem 1 are valid. Moreover, (8) and (10) hold by Lemma 1, so Theorem 1 yields the claim for X^+ . The claim for X^- is proved in a similar way. \square

4.4 Per-class bounds

In this section, we prove upper and lower bounds for the per-class number of customers in the system. Let $Z_{\lambda,\mu}^n$ be the number of customers in the

standard n -server Erlang loss system, defined as the Markov process on $\{0, 1, \dots, n\}$ having the upward transitions $x \mapsto x + 1$ at rate $\lambda 1(x < n)$ and the downward transitions $x \mapsto x - 1$ at rate μx .

Theorem 5. *Assume $Z_{\lambda_k, \mu_k}^{M_k}(0) \leq X_{1,k}(0) + X_{2,k}(0) \leq Z_{\lambda_k, \mu_k}^{M_k+N}(0)$. Then for all t ,*

$$Z_{\lambda_k, \mu_k}^{M_k}(t) \leq_{\text{st}} X_{1,k}(t) + X_{2,k}(t) \leq_{\text{st}} Z_{\lambda_k, \mu_k}^{M_k+N}(t). \quad (28)$$

Proof. Assume without loss of generality that $k = 1$. For the first inequality, construct a Markov process (\tilde{W}, \tilde{X}) on

$$S_1 = \{(w, x) \in \{0, \dots, M_1\} \times S : w \leq x_{1,1} + x_{2,1}\}$$

via the class-1 transitions ($i = 1, 2$)

$$(w, x) \mapsto (w + 1, x + e_{i,1}) \quad \text{at rate } \lambda_1 1(w < M_1, x \in A_{i,1}), \quad (29)$$

$$(w, x) \mapsto (w, x + e_{i,1}) \quad \text{at rate } \lambda_1 1(w = M_1, x \in A_{i,1}), \quad (30)$$

$$(w, x) \mapsto (w + 1, x) \quad \text{at rate } \lambda_1 1(w < M_1, x \in B_1), \quad (31)$$

$$(w, x) \mapsto (w - 1, x - e_{1,1}) \quad \text{at rate } \mu_1(w \wedge x_{1,1}), \quad (32)$$

$$(w, x) \mapsto (w - 1, x - e_{2,1}) \quad \text{at rate } \mu_1(w - x_{1,1})^+, \quad (33)$$

$$(w, x) \mapsto (w, x - e_{1,1}) \quad \text{at rate } \mu_1(x_{1,1} - w)^+, \quad (34)$$

$$(w, x) \mapsto (w, x - e_{2,1}) \quad \text{at rate } \mu_1(x_{2,1} - (w - x_{1,1}))^+, \quad (35)$$

and the class- k transitions for $k \neq 1$ and $i = 1, 2$,

$$(w, x) \mapsto (w, x + e_{i,k}) \quad \text{at rate } \lambda_k 1(x \in A_{i,k}), \quad (36)$$

$$(w, x) \mapsto (w, x - e_{i,k}) \quad \text{at rate } \mu_k x_{i,k}. \quad (37)$$

Note that all transition rates in (29) – (37) are nonnegative; for the rate in (35), observe that $w - x_{1,1} \leq x_{2,1}$ for $(w, x) \in S_1$.

Let us next verify that all transitions map S_1 into S_1 . Observe first that transition (31) occurs only if $w < M_1$ and $x_{1,1} = M_1$, which implies that $(w + 1, x) \in S_1$. Next, transition (34) occurs only if $x_{1,1} > w$, which shows that $(w, x - e_{1,1}) \in S_1$. Moreover, transition (35) occurs only if $x_{2,1} > (w - x_{1,1})^+ \geq w - x_{1,1}$, which again shows that $(w, x - e_{2,1}) \in S_1$. It is clear that all other transitions map $S_1 \rightarrow S_1$. Thus the Markov process (\tilde{W}, \tilde{X}) on S_1 is well-defined.

Moreover, the total rates of transitions in (29) – (37) where $w \mapsto w + 1$ and $w \mapsto w - 1$ are equal to $\lambda_1 1(w < M_1)$ and $\mu_1 w$, respectively. Likewise, we see that the total transition rates for $x \mapsto x + e_{i,k}$ and $x \mapsto x - e_{i,k}$ are

equal to $\lambda_k 1(x \in A_{i,k})$ and $\mu_k x_{i,k}$, respectively, for all i and k . This shows that (\tilde{W}, \tilde{X}) is a coupling of $Z_{\lambda_1, \mu_1}^{M_1}$ and X , so the first inequality in (28) is valid.

To prove the second inequality, construct a Markov process (\tilde{X}, \tilde{Y}) on

$$S_2 = \{(x, y) \in S \times \{0, \dots, M_1 + N\} : x_{1,1} + x_{2,1} \leq y\}$$

via the class-1 transitions for $i = 1, 2$,

$$(x, y) \mapsto (x + e_{i,1}, y + 1) \quad \text{at rate } \lambda_1 1(x \in A_{i,1}, y < M_1 + N), \quad (38)$$

$$(x, y) \mapsto (x + e_{i,1}, y) \quad \text{at rate } \lambda_1 1(x \in A_{i,1}, y = M_1 + N), \quad (39)$$

$$(x, y) \mapsto (x, y + 1) \quad \text{at rate } \lambda_1 1(x \in B_1, y < M_1 + N), \quad (40)$$

$$(x, y) \mapsto (x - e_{i,1}, y - 1) \quad \text{at rate } \mu_1 x_{i,1}, \quad (41)$$

$$(x, y) \mapsto (x, y - 1) \quad \text{at rate } \mu_1 (y - x_{1,1} - x_{2,1}), \quad (42)$$

and the class- k transitions for $k \neq 1$ and $i = 1, 2$,

$$(x, y) \mapsto (x + e_{i,k}, y) \quad \text{at rate } \lambda_k 1(x \in A_{i,k}), \quad (43)$$

$$(x, y) \mapsto (x - e_{i,k}, y) \quad \text{at rate } \mu_k x_{i,k}. \quad (44)$$

Note that all transition rates in (38) – (44) are nonnegative for all $(x, y) \in S_2$.

Let us now verify that all transitions map S_2 into S_2 . Observe first that transition (39) occurs only if $y = M_1 + N$ and either $x_{1,1} < M_1$ or $|x|_2 < N$, which implies that $(x + e_{i,1}, y) \in S_2$ for $i = 1, 2$. Moreover, transition (42) occurs only if $x_{1,1} + x_{2,1} < y$, so that $(x, y - 1) \in S_2$. Clearly, all other transitions map S_2 into S_2 . Thus the Markov process (\tilde{X}, \tilde{Y}) on S_2 is well-defined.

Moreover, the total rates of transitions in (38) – (44) where $x \mapsto x + e_{i,k}$ and $x \mapsto x - e_{i,k}$ are equal to $\lambda_k 1(x \in A_{i,k})$ and $\mu_k x_{i,k}$, respectively, for all i and k . The corresponding total rates for $y \mapsto y + 1$ and $y \mapsto y - 1$ are equal to $\lambda_1 1(y < M_1 + N)$ and $\mu_1 y$, respectively. This shows that (\tilde{X}, \tilde{Y}) is a coupling of X and $Z_{\lambda_1, \mu_1}^{M_1 + N}$, so the second inequality in (28) holds. \square

5 Bounds of the steady-state performance

In this section, we apply the results of Section 4 to analyze the system in steady state. We assume from now on that all arrival rates and service rates are strictly positive, which implies that all Markov processes treated in the sequel have a unique stationary distribution.

5.1 Per-class performance

Let \bar{X}_k be a random variable describing the stationary number of class- k customers in the system, and denote its mean by T_k . Moreover, denote by θ_k the stationary mean class- k throughput (the number of class- k customers completing service per unit time), and by b_k the class- k blocking probability. Note that T_k can be viewed as the mean class- k work throughput (amount of class- k work served per unit time). We use the quadruple (M, N, λ, μ) , to indicate that a performance quantity corresponds to a system with server configuration $M = (M_1, \dots, M_K)$ at layer 1, N servers at layer 2, arrival rates $\lambda = (\lambda_1, \dots, \lambda_K)$, and service rates $\mu = (\mu_1, \dots, \mu_K)$.

Let $\text{erl}(n, a)$ be a random variable on $\{0, 1, \dots, n\}$ with distribution $(\sum_{j=0}^n \frac{a^j}{j!})^{-1} \frac{a^i}{i!}$, and denote its mean by $m_{\text{erl}}(n, a)$, and the probability of being equal to n by $b_{\text{erl}}(n, a)$. Note that $b_{\text{erl}}(n, a)$ is equal to the famous Erlang B formula.

Theorem 6. *The stationary number of class- k customers in the system satisfies*

$$\text{erl}(M_k, \lambda_k/\mu_k) \leq_{\text{st}} \bar{X}_k(M, N, \lambda, \mu) \leq_{\text{st}} \text{erl}(M_k + N, \lambda_k/\mu_k). \quad (45)$$

Epecially, the stationary class- k mean number of customers is bounded by

$$m_{\text{erl}}(M_k, \lambda_k/\mu_k) \leq T_k(M, N, \lambda, \mu) \leq m_{\text{erl}}(M_k + N, \lambda_k/\mu_k), \quad (46)$$

the stationary class- k mean throughput by

$$\mu_k m_{\text{erl}}(M_k, \lambda_k/\mu_k) \leq \theta_k(M, N, \lambda, \mu) \leq \mu_k m_{\text{erl}}(M_k + N, \lambda_k/\mu_k), \quad (47)$$

and the stationary class- k blocking probability by

$$b_{\text{erl}}(M_k + N, \lambda_k/\mu_k) \leq b_k(M, N, \lambda, \mu) \leq b_{\text{erl}}(M_k, \lambda_k/\mu_k). \quad (48)$$

Proof. Let us consider a version of the queue-length process X started at $X(0) = 0$, and let $Z_{\lambda_k, \mu_k}^{M_k}$ and $Z_{\lambda_k, \mu_k}^{M_k+N}$ be as in Theorem 5, both started at zero. Because all these processes are irreducible and positive recurrent, and because stochastic ordering is closed with respect to convergence in distribution [5], (45) follows by taking $t \rightarrow \infty$ in (28).

The inequalities (46) follow by taking expectations, and the bounds (47) are a consequence of $\theta_k = \mu_k \mathbb{E} T_k$. In light of the conservation laws $\lambda_k(1 - b_k) = \theta_k$ and $\lambda_k(1 - b_{\text{erl}}) = \mu_k m_{\text{erl}}$, these bounds in turn imply (48). \square

Figure 2 illustrates the bounds in (45) for a loss network with server configuration $M = (5, 5)$ and $N = 5$, where $\lambda = (7.5, 7.5)$ and $\mu = (1, 1.3)$.

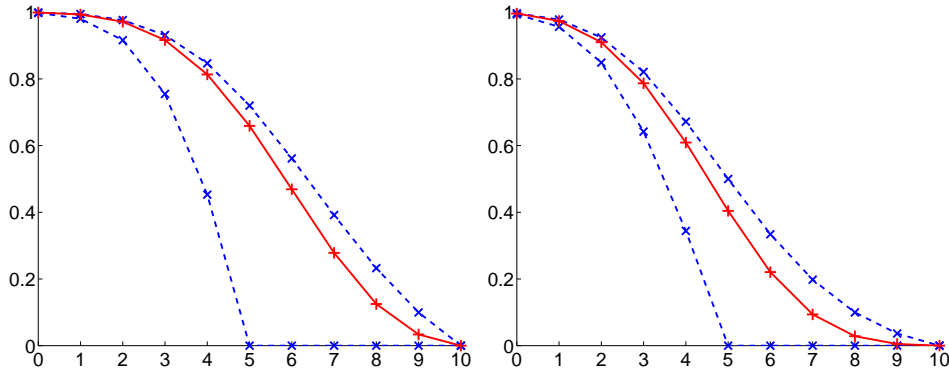


Figure 2: Complementary cumulative distribution functions of the number of customers in class 1 (left) and class 2 (right).

Remark 4. The Erlang bounds (48) for the blocking probability are well-known in the literature (see for example [1]). The stochastic inequalities (45) can be viewed as extensions of these classical bounds.

5.2 Overall performance

Let $\bar{X} = \sum_k \bar{X}_k$ be a random variable describing the stationary total number of customers in the system, and let $T = \sum_k T_k$ be its mean. Moreover, denote the stationary mean total throughput by θ , and the stationary overall blocking probability by b . Note that T may be viewed as the mean total work throughput (total amount of work served by the system in unit time). We indicate by $\bar{X}^{\text{mp}}, T^{\text{mp}}, \theta^{\text{mp}}, b^{\text{mp}}$ the corresponding quantities for a system with maximum packing.

Denote by μ^{\min} and μ^{\max} the vectors where all entries of μ are replaced by $\mu_{\min} = \min_k \mu_k$ and $\mu_{\max} = \max_k \mu_k$, respectively, and let $r_\mu = \mu_{\max}/\mu_{\min}$. Moreover, let us denote by $C_{M,N}$ the set of server configurations where all layer-2 servers have been replaced by servers in layer 1, so that

$$C_{M,N} = \{M' \in \mathbb{Z}_+^K : M'_k \geq M_k \forall k \text{ and } \sum_k M'_k = \sum_k M_k + N\}.$$

Theorem 7. *The stationary total number of customers in the system satisfies*

$$\bar{X}(M', 0, \lambda, \mu^{\max}) \leq_{\text{st}} \bar{X}(M, N, \lambda, \mu) \leq_{\text{st}} \bar{X}^{\text{mp}}(M, N, \lambda, \mu^{\min}) \quad (49)$$

for all $M' \in C_{M,N}$. Especially, the stationary mean number of customers is bounded by

$$\max_{M' \in C_{M,N}} T(M', 0, \lambda, \mu^{\max}) \leq T(M, N, \lambda, \mu) \leq T^{\text{mp}}(M, N, \lambda, \mu^{\min}), \quad (50)$$

the stationary mean throughput by

$$\max_{M' \in C_{M,N}} r_\mu^{-1} \theta(M', 0, \lambda, \mu^{\max}) \leq \theta(M, N, \lambda, \mu) \leq r_\mu \theta^{\text{mp}}(M, N, \lambda, \mu^{\min}), \quad (51)$$

and the stationary blocking probability by

$$1 - r_\mu(1 - b^{\text{mp}}(M, N, \lambda, \mu^{\min})) \leq b(M, N, \lambda, \mu) \leq \min_{M' \in C_{M,N}} (1 - r_\mu^{-1}(1 - b(M', 0, \lambda, \mu^{\max}))). \quad (52)$$

Remark 5. In the case where all service rates μ_k are equal, the bounds (51) and (52) can be written in a more natural form as

$$\max_{M' \in C_{M,N}} \theta(M', 0, \lambda, \mu) \leq \theta(M, N, \lambda, \mu) \leq \theta^{\text{mp}}(M, N, \lambda, \mu),$$

and

$$b^{\text{mp}}(M, N, \lambda, \mu) \leq b(M, N, \lambda, \mu) \leq \min_{M' \in C_{M,N}} b(M', 0, \lambda, \mu).$$

Remark 6. The upper and lower bounds in (49), and hence the also the bounds in (50) – (52), are easy to compute numerically. The fast computation of the upper bound is explained in Remark 1. To compute the lower bound, observe that $\bar{X}(M', 0, \lambda, \mu^{\max})$ has the same distribution as $\sum_k \text{erl}(M'_k, \lambda_k / \mu_{\max})$, where the terms in the sum are independent.

Proof of Theorem 7. Let X be the queue-length process of the original system, let W be the queue-length process in the system corresponding to the parameters $(M', 0, \lambda, \mu^{\max})$, and let Y be the queue-length process of the maximum packing system with parameters $(M, N, \lambda, \mu^{\min})$. Assume that all processes are started at zero initial state. Then Theorem 2 and Theorem 3 combined with Theorem 4 imply that

$$|W(t)| \leq_{\text{st}} |X(t)| \leq_{\text{st}} |Y(t)| \quad (53)$$

for all t . Because all of the above processes are irreducible and positive recurrent, and because stochastic ordering is closed with respect to convergence in distribution [5], taking $t \rightarrow \infty$ in (53) shows the validity of (49). The bounds in (50) follow by taking expectations, and the bounds in (51) from $\theta = \sum_k \mu_k T_k$. These bounds in turn imply (52), because of the conservation law $(\sum_k \lambda_k)(1 - b) = \theta$. \square

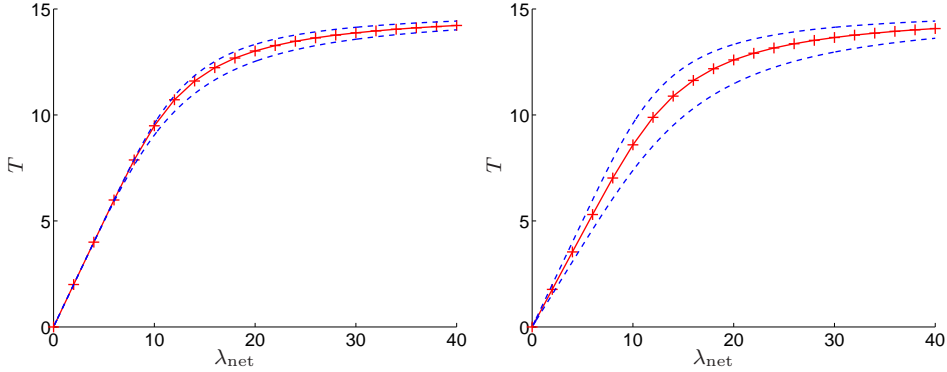


Figure 3: Bounds of the mean number of customers for $M = (5, 5)$, $N = 5$, and $\lambda = (\lambda_{\text{net}}/2, \lambda_{\text{net}}/2)$; where $\mu = (1, 1)$ on the left, and $\mu = (1, 1.3)$ on the right.

Figure 3 illustrates the bounds (50) of the mean number of customers in a two-class system where the net arrival rate λ_{net} is varying. The right plot shows that the bounds become slightly less accurate when the service rates are different. This lack of accuracy is mainly a consequence of replacing μ by μ^{\min} and μ^{\max} in (49).

Intuitively one might think that the upper bound in (49) might be strengthened to hold without replacing μ by μ^{\min} . Example 1 shows that this is not true in general. A similar observation for a model describing the assignment of channels in cellular radio networks has been made by Kelly [6].

Example 1. Consider a two-class loss network with server configuration $M = (1, 0)$ and $N = 2$. Assume $\lambda = (1, 1)$ and $\mu = (\frac{1}{5}, 10)$, so that the service rates differ from each other by a factor of 50. Table 1 lists numerically calculated values of the stationary mean number of customers (per class and total) for the original loss network and the modification with maximum packing.

	Class 1	Class 2	Total
$T(M, N, \lambda, \mu)$	2.325657	0.038612	2.364269
$T^{\text{mp}}(M, N, \lambda, \mu)$	2.317818	0.046344	2.364162
$T(M, N, \lambda, \mu^{\min})$	1.615744	0.997537	2.613281
$T^{\text{mp}}(M, N, \lambda, \mu^{\min})$	1.474617	1.172442	2.647059

Table 1: Mean number of customers in a loss network with and without maximum packing.

Example 2 shows that replacing one layer-2 server by a layer-1 server may not decrease the stationary mean number of customers, if not all service rates μ_k are equal. This shows that it is necessary to replace μ by μ^{\max} in order to achieve a lower bound in (49).

Example 2. Consider a two-class loss network with two different server configurations (i) $M = (0, 0)$ and $N = 3$, and (ii) $M' = (1, 0)$, $N' = 2$. Assume that λ and μ are as in Example 1. Numerically calculated values for the stationary mean number of customers (per class and total) given in Table 2.

	Class 1	Class 2	Total
$T(M, N, \lambda, \mu)$	2.317808	0.046356	2.364164
$T(M', N', \lambda, \mu)$	2.325657	0.038612	2.364269
$T(M, N, \lambda, \mu^{\max})$	0.099891	0.099891	0.199782
$T(M', N', \lambda, \mu^{\max})$	0.099906	0.099453	0.199359

Table 2: Mean number of customers in a loss network with two different server configurations.

6 Conclusions

Stochastic comparison and coupling techniques were developed for analyzing multiclass two-layer loss systems. First, assuming all service rates to be equal, we proved that maximum packing stochastically increases the total number of customers, and that moving a server from the second layer to the first has the opposite effect. The monotonicity of the system with respect to service rates was then shown to extend the above conclusions to systems where the service rates may differ from each other. As a consequence, easily computable upper and lower bounds for the performance of the system were derived. An important topic for future research is to develop analytical methods for quantifying the tightness of these bounds.

Acknowledgments

We gracefully acknowledge helpful discussions with Sem Borst. Part of this research has been funded by the Dutch BSIK/BRICKS PDC2.1 project.

References

- [1] Borst, S. and Whiting, P. A. (2000). Achievable performance of dynamic channel assignment schemes under varying reuse constraints. *IEEE T. Veh. Technol.*, 49(4):1248–1264.
- [2] Everitt, D. E. and Macfadyen, N. W. (1983). Analysis of multicellular mobile radiotelephone systems with loss. *Brit. Telecom Technol. J.*, 1(2):37–45.
- [3] Franx, G. J., Koole, G., and Pot, A. (2006). Approximating multi-skill blocking systems by hyperexponential decomposition. *Perform. Evaluation*, 630:799–824.
- [4] Hordijk, A. and Ridder, A. (1987). Stochastic inequalities for an overflow model. *J. Appl. Probab.*, 24:696–708.
- [5] Kamae, T., Krengel, U., and O’Brien, G. L. (1977). Stochastic inequalities on partially ordered spaces. *Ann. Probab.*, 5(6):899–912.
- [6] Kelly, F. P. (1985). Stochastic models of computer communication systems. *J. Roy. Stat. Soc. B*, 47(3):379–395.
- [7] Kelly, F. P. (1991). Loss networks. *Ann. Appl. Probab.*, 1:319–378.
- [8] Louth, G., Mitzenmacher, M., and Kelly, F. P. (1994). Computational complexity of loss networks. *Theor. Comp. Sc.*, 125:45–59.
- [9] Massey, W. A. (1987). Stochastic orderings for Markov processes on partially ordered spaces. *Math. Oper. Res.*, 12(2):350–367.
- [10] Müller, A. and Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*. Wiley.
- [11] Nain, P. (1990). Qualitative properties of the Erlang blocking model with heterogeneous user requirements. *Queueing Syst.*, 6:189–206.
- [12] Thorisson, H. (2000). *Coupling, Stationarity, and Regeneration*. Springer.
- [13] Whitt, W. (1981). Comparing counting processes and queues. *Adv. Appl. Probab.*, 13(1):207–220.
- [14] Wolff, R. W. (1989). *Stochastic Modeling and the Theory of Queues*. Prentice Hall.