Bottleneck Analysis in Multiclass Closed Queueing Networks and Its Application

ARTHUR BERGER[1], LEV BREGMAN[2] AND YAAKOV KOGAN[3]

[1] *Bell Labs, Lucent Technologies, Holmdel, NJ 07733, USA*
E-mail: awberger@lucent.com

[2] *Institute of Industrial Mathematics, Beer-Sheva, Israel*
E-mail: bregman@bengus.bgu.ac.il

[3] *AT&T Labs, Holmdel, NJ 07733, USA*
E-mail: yakov@buckaroo.att.com

Asymptotic behavior of queues is studied for large closed multi-class queueing networks consisting of one infinite server station with K classes and M processor sharing (PS) stations. A simple numerical procedure is derived that allows us to identify all bottleneck PS stations. The bottleneck station is defined asymptotically as the station, where the number of customers grows proportionally to the total number of customers in the network as the latter increases simultaneously with service rates at PS stations. For the case when $K = M = 2$, the set of network parameters is identified that corresponds to each of the three possible types of behavior in heavy traffic: both PS stations are bottlenecks, only one PS station is a bottleneck and a group of two PS stations is a bottleneck while neither PS station forms a bottleneck by itself. In the last case both PS stations are equally loaded by each customer class and their individual queue lengths, normalized by the large parameter, converge to uniformly distributed random variables. These results are directly generalized for arbitrary $K = M$. Generalizations for $K \neq M$ are also indicated. The case of two bottlenecks is illustrated by its application to the problem of dimensioning bandwidth for different data sources in packet-switched communication networks. An engineering rule is provided for determining the link rates such that a service objective on a per-class throughput is satisfied.

Keywords: Closed queueing networks, asymptotic analysis, bottleneck.

1 Introduction

This paper is motivated by a new application of closed queueing networks (CQN) with a large number of customers. The application is the dimensioning bandwidth for different data sources subject to feedback control in packet-switched communication networks when available bandwidth at the servers is shared between all active sources. In a CQN, data sources are modeled by an infinite server (IS) station, and network nodes are modeled by processor sharing (PS) stations. It is known that the steady state queue length distribution in such a CQN has a product form that is defined explicitly up to the normalization constant. The distinguishing property of the new application is that this CQN model is adequate only if one or more PS stations form a bottleneck. The bottleneck station is defined asymptotically as the station, where the number of customers grows proportionally to the total number of customers in the network as the latter increases simultaneously with service rates at PS stations. It is known [7] that for a single class CQN ($K = 1$) consisting of IS and PS stations, in general, only one PS station may be a bottleneck, and the bottleneck node can be easily identified from the network parameters. Moreover, the asymptotics for the mean queue length at the bottleneck station is found from a linear equation. For a multiclass CQN the bottleneck analysis becomes more complicated. For an arbitrary number of classes, the bottleneck analysis has been done only in the case when the number of PS stations $M = 1$ [9]. In this case, the asymptotics of the mean queue length at the bottleneck node are explicitly expressed through the least positive root of a polynomial of order K, where K is the number of classes. The relative simplicity of the results for $K = 1,\ M > 1$ or $K > 1,\ M = 1$ is explained by their derivation from asymptotic expansions of one-dimensional integral representations for the partition function (normalization constant) in complex [2, 6, 7] or real [9] space. In general, the integral representations in complex and real space are K- and M-dimensional respectively, and their asymptotics can be relatively easy derived only in the case of normal traffic [9, 7] when neither a PS station nor or a group of PS stations forms a bottleneck. The bottleneck case requires residue analysis of a K-dimensional generating partition function and application of the saddle-point method, which is far from trivial even in the two-dimensional case [8], or non-traditional application of the M-dimensional Laplace method that, to our knowledge, has not been pursued. Moreover, in all cases but one, the bottleneck conditions given in [8] are quite complicated, their probabilistic interpretation and generalization for $K,\ M > 2$ is unclear, and the case of equally loaded PS stations is not covered.

Therefore, in this paper, we take a direct approach based on the asymptotic representation for the steady state queue length distribution $\pi(\mathbf{n})$ derived by Pittel [11]. This representation has the following form:

$$\pi(\mathbf{n}) \sim C \exp\{NF(\mathbf{x})\},$$

where N is a large parameter (e.g., the total number of customers) and $\mathbf{x} = \mathbf{n}/N$. Pittel showed that bottleneck nodes in a large product-form CQN can be identified by a non-zero maximum point of some multidimensional function $F(\mathbf{x})$ under natural constraints. Positive components of the maximum point $\mathbf{x}^*$ are the limiting values of queue lengths, normalized by N, at the bottleneck nodes.

Pittel derived these results under the condition that the optimization problem has a unique solution. The questions, how is this condition expressed in terms of the network parameters, what is the range of network parameters for which the maximum point is not zero and how to solve the optimization problem, were not addressed.

In our case, the $(K \times M)$-dimensional function F is found explicitly. We show that all possible maxima of F under naturally defined constraints can be efficiently found by the classical Lagrange Multiplier Method. Using this method we show that maximization of F can be reduced to that for an M-dimensional function. Moreover, for this M-dimensional function we derive a simple formula for its partial derivatives that plays a pivotal role in the bottleneck identification depending on the network parameters. Finally, the calculation of $\mathbf{x}^*$ is reduced to the solution of algebraic equations and verification of inequalities. The number of equations and inequalities equals the number of bottleneck and non-bottleneck nodes respectively. In general, these equations are nonlinear. But in an important case when $K \leq M$ and K bottlenecks, $\mathbf{x}^*$ can be found by solving two systems of linear equations of order K.

Note that efficient calculation of $\mathbf{x}^*$ is also important for the computation of the normalization constant in the initial product-form solution for a large network with bottlenecks. This is because by its construction the function $F(\mathbf{x})$ is a quasi-potential [4] that provides the logarithmic asymptotic for the product-form solution, and hence $\exp\{NF(\mathbf{x}^*)\}$ can be used as a scaling factor in the computation of the normalization constant.

The outline of the paper is as follows. In section 2 we describe the closed queueing network, define the scaling under which we study the asymptotics of the steady state distribution and provide the expression for the function F. In section 3 we formulate the main results in two theorems. The first theorem addresses a special case of the normalized queue-length limit behavior which is referred to as oscillation. The second theorem provides bottleneck classification for all possible combinations of the network parameters. We consider the case $K = M = 2$ but whenever it is possible the results are formulated in a general form. In section 4 we illustrate the case of two bottlenecks by its application to the problem of dimensioning bandwidth for elastic data sources in packet-switched communication networks. In section 5 we, first, establish in three lemmas important properties of the maximum of the function F and then use them to prove the theorems. In section 6, we indicate generalizations for arbitrary K and M.

2 Asymptotic representation for the steady state distribution

We consider a closed queueing network with K classes and $M+1$ service stations, one of which is infinite server (IS) and M others are processor-sharing (PS) stations. We assume that customers of each class visit all stations. It is convenient to number the IS-station by 0. Let $\mathbf{n}$ denote a $K \times M$ matrix whose element n_{ki} represents the number of class k customers at PS-station i. The population of jobs in class k is a constant N_k, $1 \le k \le K$. The state space is the set $\mathcal{S}$ of matrices $\mathbf{n}$ which have integer components, and satisfy the population constraints

$$\mathcal{S} = \{\mathbf{n} | 0 \le n_{ki},\ \sum_i n_{ki} \le N_k,\ \ 1 \le k \le K,\ \ 1 \le i \le M\}.$$

Then the product form solution has the form:

$$\pi(\mathbf{n}) = \frac{1}{\Omega} \prod_{k=1}^{K} \frac{N_k!}{(N_k - \sum_i n_{ki})!} \prod_{i=1}^{M} n_i! \frac{r_{ki}^{n_{ki}}}{n_{ki}!}, \tag{1}$$

where Ω is the normalization constant and $n_i = \sum_k n_{ki}$. Moreover,

$$r_{ki} = \frac{e_{ki}\lambda_k}{\mu_{ki}}, \tag{2}$$

where e_{ki} is the relative visiting rate of class k jobs to PS-station i as compared to the IS-station, $1/\lambda_k$ is the mean service time of a class k job at the IS-station, $1/\mu_{ki}$ is the mean service time of an isolated class k job at PS-station i.

Denote by Q_{ki} the random variable for the number of k-type customers in service (queue length) at PS-station i and by $\mathbf{Q}$ the $K \times M$ matrix of these queue lengths. Random matrix $\mathbf{Q}$ takes values $\mathbf{n} \in \mathcal{S}$. Our goal is to study limit behavior of $\mathbf{Q}$ under the following assumption:

$$\alpha_k = N_k/N,\ \ \rho_{ki} = N r_{ki}, \tag{3}$$

where α_k and ρ_{ki} are positive constants while $N \to \infty$. (Note that this scaling is reasonable for the intended application to data networks, section 4, where large values of N correspond to the important case of a large number of established connections (or flows) and to high-speed transmission facilities, i.e., large values of μ_{ki} and small r_{ki}.) With this assumption we have the following asymptotic representation [11]

$$\pi(\mathbf{n}) = C(N) \exp\{N F(\mathbf{x}) + O(\ln N)\} \tag{4}$$

with

$$F(\mathbf{x}) = \sum_i (x_i \ln x_i - x_i) + \sum_{k,i} x_{ki} \ln \rho_{ki} - \sum_{k,i} x_{ki} \ln x_{ki} - \sum_k (\alpha_k - \sum_i x_{ki}) \ln(\alpha_k - \sum_i x_{ki}), \tag{5}$$

where $x_{ki} = n_{ki}/N$, $x_i = \sum_k x_{ki}$ for $1 \le k \le K$, $1 \le i, \le M$, and $C(N)$ does not depend on $\mathbf{x}$. ¿From the definition of the variables x_{ki} it follows that $\mathbf{x} \in \mathcal{C}$, where

$$\mathcal{C} = \{\mathbf{x} : x_{ki} \ge 0 \text{ and } \sum_i x_{ki} \le \alpha_k\}. \tag{6}$$

3 Asymptotic behavior of queues at PS-stations

In this section we state the asymptotic results in two theorems and briefly comment on them, deferring the proofs to section 5. We consider the case $K = M = 2$ but whenever it is possible the results are formulated in a general form. In section 6 we discuss generalizations of the results for different cases when $K,\ M \ge 2$.

Pittel [11] assumed that the function $F(\mathbf{x})$ has a unique maximum $\mathbf{x}^*$ and proved the convergence in probability of the normalized queue length matrix $\mathbf{Q}/N$ to $\mathbf{x}^*$ as $N \to \infty$. The first theorem characterizes the limit behavior of $\mathbf{Q}/N$ in a special case of network parameters where it does not converge to a deterministic limit.

Theorem 1 Oscillation
If $\rho_{ki} \equiv \rho_k$ and $\sum_k \rho_k \alpha_k > 1$ then

$$(1/N) \sum_{k,i} Q_{ki} \xrightarrow{P} v^* \tag{7}$$

as $N \to \infty$, where v^* is the unique solution of equation

$$\sum_k \frac{\rho_k \alpha_k}{\rho_k v + 1} = 1 \tag{8}$$

in the interval $(0, \sum_k \alpha_k)$. Moreover,

$$\frac{Q_{11} + Q_{12}}{N} \xrightarrow{P} u^*, \tag{9}$$

where

$$u^* = \alpha_1 - \frac{\alpha_1}{\rho_1 v^* + 1}, \tag{10}$$

and

$$\frac{Q_{11}}{N} \xrightarrow{D} U(u^*), \tag{11}$$

while

$$\frac{Q_{21}}{N} \xrightarrow{D} U(z^*), \tag{12}$$

as $N \to \infty$, where $U(d)$ denotes a random variable with the uniform distribution on $[0, d]$, and $z^* = v^* - u^*$.

The type of the queue length limit behavior at an individual PS station described by (11) or (12) is referred to as *oscillation*.

In section 5 we show that the function $F(\mathbf{x})$ has a unique maximum $\mathbf{x}^*$ except of the special case of network parameters in theorem 1. Moreover,

$$x_i^* = \sum_k x_{ki}^* > 0$$

if and only if $x_{ki}^* > 0$ for each class k. This property in combination with the convergence $\mathbf{Q}/N \xrightarrow{P} \mathbf{x}^*$, [11], justifies the following definition.

A PS-station i is referred to as *bottleneck* (*non-bottleneck*) if $x_i^* > 0$ ($x_i^* = 0$), given that the maximum point $\mathbf{x}^*$ is unique. Statement (7) addresses a special case, where the bottleneck condition is satisfied only for a group of stations. We say that a group $B = (i_1, \ldots, i_m)$ of $m \leq M$ PS stations forms a bottleneck if they are equally loaded, i.e. $\rho_{ki} \equiv \rho_k(B),\ i \in B$ and the following normalized sum of queue lengths

$$(1/N)\sum_{i\in B}\sum_k Q_{ki}$$

converges to a deterministic limit $v^*(B) > 0$ in probability as $N \to \infty$.

We say that a PS-station (a group of equally loaded PS-stations) is *heavy loaded* if it forms a bottleneck.

Representation (4) implies that distribution (1) is concentrated around maximum points of the function $F(\mathbf{x})$ in the domain $\mathcal{C}$. It turns out that the function F does not have the unique maximum only in the case when PS stations are equally heavy loaded by each class of customers. This results in random "oscillation" of the individual queue lengths, normalized by N, in contrast to their stabilization to deterministic limits in the cases of the unique maximum. Load balancing is a plausible design decision in many applications. However, except for a simple cyclic network consisting of identical FCFS single servers [5], it was not clear before that although load balancing indeed equalizes the mean queue lengths at different nodes the actual normalized queue lengths are uniformly distributed random variables.

Denote by $\Delta = \rho_{11}\rho_{22} - \rho_{12}\rho_{21}$ the determinant of matrix $\|\rho_{ki}\|$. The next theorem considers all possible combinations of parameters ρ_{ki} and α_k and provides a complete bottleneck classification in the case $K = M = 2$.

Theorem 2 Bottleneck classification

1. *If* $\sum_k \rho_{ki}\alpha_k > 1$ *for all* i, $\Delta \neq 0$ *and the two following systems*

$$\sum_k \rho_{ki}\beta_k = 1,\ \ i = 1, 2, \tag{13}$$

$$\sum_i \rho_{ki}\gamma_i = (\alpha_k - \beta_k)/\beta_k,\ \ k = 1, 2 \tag{14}$$

have positive solutions, then

$$x^*_{ki} = \rho_{ki}\beta_k\gamma_i, \tag{15}$$

and all PS-stations are bottlenecks.

2. *If* $\sum_k \rho_{ki}\alpha_k > 1$ *for* $i = 1,\ 2$, $\Delta = 0$ *and equations (13) have a solution* $\beta_k \in (0, \alpha_k)$, *then* $\rho_{ki} \equiv \rho_k$ *and the group of two PS-stations forms a bottleneck with oscillation at the individual PS-stations.*

3. *In the two following cases there is only one bottleneck PS-station.*

A. $\sum_k \rho_{ki}\alpha_k > 1$ *only for one* i.

B. $\sum_k \rho_{ki}\alpha_k > 1$ *for* $i = 1,\ 2$ *but equation (13) does not have a solution* $\beta_k \in (0, \alpha_k)$ *or* $\Delta \neq 0$ *and (14) does not have a positive solution.*

In case **A** *PS-station* i *is the bottleneck. In case* **B** *PS-station* i *is the bottleneck if*

$$\sum_k \frac{\rho_{kj}\alpha_k}{\rho_{ki}\gamma^*_i + 1} < 1 \tag{16}$$

for $j \neq i$, *where* γ^*_i *is the unique positive solution of equation*

$$\sum_k \frac{\rho_{ki}\alpha_k}{\rho_{ki}\gamma + 1} = 1. \tag{17}$$

If PS-station i *is the only bottleneck, then*

$$x^*_{ki} = \frac{\rho_{ki}\alpha_k}{\rho_{ki} + 1/\gamma^*_i}, \tag{18}$$

where γ^*_i *is the unique positive solution of equation (17).*

4. *If* $\sum_k \rho_{ki}\alpha_k \leq 1$ *for all* i, *then* $\mathbf{x}^* = \mathbf{0}$, *and all PS-station are non-bottleneck.*

Condition (16) has the following interpretation. By (18) the left hand side of the inequality (16) coincides with the sum of traffic intensities in an open system in which PS-station j serves two types of Poissonian arrivals whose rates are $e_{kj}\lambda_k N(\alpha_k - x^*_{ki})$, $k = 1,\ 2$. Thus (16) means that the sum of traffic intensities is less than 1 at a non-bottleneck PS-station, i.e. the open system is stable.

We use the classical Lagrange Multiplier Method to find all possible local maxima of the function $F(\mathbf{x})$ in the domain $\mathcal{C}$. Using this method we derive in section 5 the necessary conditions for a maximum and reduce the initial problem to maximization of a function of M variables $x_1, \ldots, x_M$. It turns out that the latter function depends only on $\sum_{i\in B} x_i$ if a group of nodes B forms a bottleneck.

4 Bandwidth dimensioning for elastic data sources

In this section we illustrate statement 1 of theorem 2 by its application to the problem of dimensioning bandwidth for different data sources in packet-switched communication networks, such as Internet Protocol (IP) or Asynchronous Transfer Mode (ATM) networks. For further details see [1]. In our application, a type-k

job is a file with mean size of f_k bits, and the link rate at server i is L_i bps (bits per second). Then the service rate of a type-k job at node i (given no other job is present) is

$$\mu_{ki} = L_i / f_k. \tag{19}$$

Suppose, for simplicity, that there are two job types generated by finite sources and two bottleneck links. This mirrors the important case in data networks under heavy load, where a routing algorithm such as Private Network–Network Interface (P-NNI), [12] directs the traffic to otherwise lightly loaded paths. We model finite data sources and network nodes by IS and PS stations respectively and obtain a CQN model with $K = M = 2$. Let, also for simplicity,

$$e_{11} = p,\ e_{12} = 1 - p,\ e_{21} = q,\ e_{22} = 1 - q,\quad 0 \leq p, q \leq 1. \tag{20}$$

We assume that for a given set of parameters $\{N_k,\ \lambda_k,\ f_k,\ p,\ q\}$ the link rates L_i are such that the conditions of statement 1 in theorem 2 are satisfied, i.e., both PS-stations are bottlenecks. Thus we can approximate the sum of the throughputs of type-k jobs at two nodes by

$$T_k = \frac{\mu_{k1} x^*_{k1}}{x^*_{11} + x^*_{21}} + \frac{\mu_{k2} x^*_{k2}}{x^*_{12} + x^*_{22}}, \tag{21}$$

where x^*_{ki} are given by (15). Substituting into (21) the values of μ_{ki} and x^*_{ki} from (19) and (15) respectively we get

$$T_k f_k = \frac{L_1 \rho_{k1} \beta_k}{\rho_{11}\beta_1 + \rho_{21}\beta_2} + \frac{L_2 \rho_{k2} \beta_k}{\rho_{12}\beta_1 + \rho_{22}\beta_2} = \beta_k (L_1 \rho_{k1} + L_2 \rho_{k2}), \tag{22}$$

where the last equality is implied by (13). If we sum the per-class throughputs, k=1,2, in (22) and again apply (13) we obtain:

$$T_1 f_1 + T_2 f_2 = L_1 + L_2. \tag{23}$$

(23) is the intuitively clear statement that when both PS stations are bottlenecks, the total throughput (the sum of the per-class throughputs in bps) equals the total capacity. Thus, given the conditions of statement 1 in theorem 2, from the viewpoint of total throughput it does not matter what are the individual values of the link capacities, only their sum. Likewise, note that (23) does not depend on the particulars of the routing, the e_{ki}'s, other than that statement 1 of theorem 2 pertains. This prediction from the model matches the rule of thumb in data networks that for given deployed capacity, the traffic will find the spare bandwidth via the adaptive routing.

Suppose a network designer wants to dimension the capacity of links $i = 1, 2$ to provide a service objective based on throughput. If the chosen objective is in terms of the total throughput for both classes, then from (23) the total bandwidth, $L_1 + L_2$ simply needs to be equal to the objective on total throughput. If, however, the network designer wishes to offer an objective on per-class throughput, say the bandwidth provided to type-k jobs should be at least M_k bps, then L_1 and L_2 need to be chosen such that

$$T_k f_k \geq M_k, \quad k = 1, 2. \tag{24}$$

The bandwidth dimensioning problem consists of determination of link rates L_1 and L_2 that guarantee the service objective.

Proposition 3
If

$$L_1 = \kappa[pM_1 + qM_2] \tag{25}$$

$$L_2 = \kappa[(1-p)M_1 + (1-q)M_2], \tag{26}$$

where $\kappa \geq 1$, and if the conditions of statement 1 in theorem 2 pertain, then the per-class throughput objective (24) is satisfied.

Note that for the per-class throughput objective (24), the dimensioned bandwidth only depends on the given objective M_k and the routing via p and q, and not on any other system parameters, other than via the conditions of statement 1 in theorem 2.

Proof
Substituting (19) in (2) and using (3) we have

$$\rho_{ki} = N\lambda_k f_k e_{ki} / L_i. \tag{27}$$

Substituting (27) and (20) into (22) yields

$$\theta_k \equiv T_k f_k = N\lambda_k f_k \beta_k. \tag{28}$$

One of the conditions of statement 1 in theorem 2 is that $\Delta \equiv \rho_{11}\rho_{22} - \rho_{12}\rho_{21} \neq 0$. ¿From (27) and (20)

$$\Delta = \frac{N^2\lambda_1\lambda_2 f_1 f_2}{L_1 L_2}(e_{11}e_{22} - e_{12}e_{21}) = \frac{N^2\lambda_1\lambda_2 f_1 f_2}{L_1 L_2}(p-q). \tag{29}$$

Thus we require $p \neq q$. Another condition of statement 1 in theorem 2 is that the system (13) has a positive solution β_k. For ρ_{ki} in (27), this condition implies:

$$\beta_1 = \frac{1}{N\lambda_1 f_1}\frac{(1-q)L_1 - qL_2}{p-q} > 0 \tag{30}$$

$$\beta_2 = \frac{1}{N\lambda_2 f_2}\frac{pL_2 - (1-p)L_1}{p-q} > 0. \tag{31}$$

Substituting $\beta_k,\ k=1,2$ from (30) and (31) into (28) yields

$$\theta_1 = \frac{(1-q)L_1 - qL_2}{p-q} \tag{32}$$

$$\theta_2 = \frac{pL_2 - (1-p)L_1}{p-q}, \tag{33}$$

where $\theta_k > 0,\ k=1,2$. Substituting (25) and (26) in (32) and (33) we get

$$\theta_k = \kappa M_k, \quad k=1,2$$

and condition (24) is implied by the definition of θ_k in (28) since $\kappa \geq 1$.

5 Proofs

In this section we, first, establish in three lemmas the properties of the maximum of the function $F(\mathbf{x})$. Then we proof the two theorems using these properties.

5.1 Properties of the maximum

Lemma 4
If $F(\mathbf{x})$ has a maximum at point $\mathbf{x}^*$ with $x^*_{ki} > 0\ k,i=1,2$, then the system of linear equations (13) has such a solution $\beta_k \in (0,\alpha_k),\ k=1,2$ that the system of linear equations (14) has a solution $\gamma_i > 0,\ i=1,2$. Moreover, $x^*_{ki} = \rho_{ki}\beta_k\gamma_i$.

Proof
Without the loss of generality we can assume that $\sum_i x^*_{ki} < \alpha_k,\ k=1,2$ since $\partial F(\mathbf{x})/\partial x_{ki} \rightarrow -\infty$ as $\sum_i x_{ki} \rightarrow \alpha_k,\ k=1,2$.

Denote $x=(x_1,x_2)$, $y=(y_1,y_2)$ and consider the following auxiliary problem: Maximize

$$H(x,y,\mathbf{x}) = \sum_i (x_i \ln x_i - x_i) + \sum_{k,i} x_{ki}\ln\rho_{ki} - \sum_{k,i} x_{ki}\ln x_{ki} - \sum_k y_k \ln y_k \tag{34}$$

subject to constraints

$$\sum_k x_{ki} = x_i,\ \sum_i x_{ki} + y_k = \alpha_k,\ k,i=1,2 \tag{35}$$

on the set

$$x_{ki} > 0,\ y_k > 0,\ k,i=1,2. \tag{36}$$

It is clear that $\mathbf{x}^*$ provides a solution for the auxiliary optimization problem. On the other hand, the classical Lagrange Multiplier Method provides the following necessary conditions for a local maximizer for the auxiliary problem [10, Theorem 7.2.1]. There exist $(\gamma^1, \gamma^2, \beta^1, \beta^2)$ such that:

$$\nabla H(x^*, y^*, \mathbf{x}^*) + \sum_i \gamma_i^1 \nabla(\sum_k x_{ki}^* - x_i^*) + \sum_k \beta_k^1 \nabla(\sum_i x_{ki}^* + y_k^* - \alpha_k) = 0.$$

or

$$\ln x_i^* - \ln \gamma_i = 0, \; i = 1, 2,$$
$$-\ln y_k^* + \ln \beta_k = 0, \; k = 1, 2,$$
$$\ln \rho_{ki} - \ln x_{ki}^* + \ln \gamma_i + \ln \beta_k = 0, \; i, k = 1, 2,$$

where

$$\gamma_i = \exp(\gamma_i^1), \quad \beta_k = \exp(\beta_k^1 - 1).$$

Thus we have

$$\gamma_i = x_i^*, \; \beta_k = y_k^*, \; x_{ki}^* = \rho_{ki}\gamma_i\beta_k, k = 1, 2. \tag{37}$$

We have from (37) and the definition of x_i that (β_1, β_2) satisfy the system (13). Furthermore, (37) and condition $\sum_i x_{ki}^* + y_k^* = \alpha_k$ imply that (γ_1, γ_2) satisfy the system (14). Now, to complete the proof we note that γ_i and β_k satisfy the required constraints by their definition in (37).

Lemma 5
Let $x = (x_1, x_2)$, $y = (y_1, y_2)$, and

$$S = \{x : x_1 \geq 0, \; x_2 \geq 0, \; \sum_i x_i \leq \sum_k \alpha_k\}.$$

and for any $x \in S$

$$D(x) = \{(y, \mathbf{x}) : \sum_k x_{ki} = x_i, \; \sum_i x_{ki} + y_k = \alpha_k, \; y_k \geq 0, \; x_{ki} \geq 0, \quad k, i = 1, 2\},$$

Let

$$G(y, \mathbf{x}) = \sum_{k,i} x_{ki} \ln \rho_{ki} - \sum_{k,i} x_{ki} \ln x_{ki} - \sum_k y_k \ln y_k,$$
$$K(x) = \max_{(y,\mathbf{x}) \in D(x)} G(y, \mathbf{x}),$$

and

$$R(x) = \sum_i (x_i \ln x_i - x_i) + K(x).$$

Then

$$\max_{\mathbf{x} \in \mathcal{C}} F(\mathbf{x}) = \max_{x \in S} R(x). \tag{38}$$

Moreover, let

$$x_i > 0 \text{ and } \sum_i x_i < \sum_k \alpha_k,\ i = 1, 2. \tag{39}$$

Then a maximum point $(y^0, \mathbf{x}^0)$ *of* $G(y, \mathbf{x})$ *in* $D(x)$ *satisfies*

$$y_i^0 > 0,\ x_{ki}^0 > 0,\ k, i = 1, 2$$

and

$$\frac{\partial R(x)}{\partial x_i} = \ln \sum_k \rho_{ki} y_k^0. \tag{40}$$

Proof

$G(y, \mathbf{x})$ is a strictly concave function on convex set $D(x)$, and this set has an interior point in non-degenerate case when at least one of $x_i > 0$. Therefore, for any fixed $x \in S$, $G(y, \mathbf{x})$ has the unique maximum, and function $K(x)$ is well defined. Now, (38) follows from the definition of $R(x)$.

Under conditions (39) the function G cannot have a maximum in $D(x)$ when one of $x_{ki} = 0$ or $y_k = 0$ since partial derivatives of G with respect to x_{ki} and y_k tend to ∞ as the respective variable approaches 0. The Lagrangian for the maximization problem of G under constraints (35) on the set (36) is the function

$$L(y, \mathbf{x}, \gamma^1, \gamma^2, \beta^1, \beta^2) = G(y, \mathbf{x}) + \sum_i \gamma^i (\sum_k x_{ki} - x_i) + \sum_k \beta^k (\sum_i x_{ki} + y_k - \alpha_k).$$

By Kuhn-Tucker theorem there is a vector $(\gamma_0^1, \gamma_0^2, \beta_0^1, \beta_0^2)$ such that

$$G(y^0, \mathbf{x}^0) = \max_{x_{ki} > 0, y_k > 0} L(y, \mathbf{x}, \gamma_0^1, \gamma_0^2, \beta_0^1, \beta_0^2) \tag{41}$$

The necessary conditions for a local maximum of L imply

$$-\ln y_k^0 + \ln \beta_k^0 = 0,\ k = 1, 2,$$

$$\ln \rho_{ki} - \ln x_{ki}^0 + \ln \gamma_i^0 + \ln \beta_k^0 = 0,\ i, k = 1, 2,$$

where

$$\gamma_i^0 = \exp(\gamma_0^i),\quad \beta_k^0 = \exp(\beta_0^k - 1).$$

Thus we have

$$\beta_k^0 = y_k^0,\ x_{ki}^0 = \rho_{ki} \gamma_i^0 \beta_k^0, k = 1, 2. \tag{42}$$

¿From (42) and the definition of x_i we have

$$\gamma_0^i = \ln x_i - \ln \sum_k \rho_{ki} y_k^0.$$

Hence (40) follows from the definition of $R(x)$ and (41) since $\partial L / \partial x_i = -\gamma^i$.

Lemma 6
If $\Delta \neq 0$ and $R(x)$ has a maximum on the axis $x_{3-i} = 0$, then $\gamma_{3-i} < 0,\ i = 1,\ 2$, where $(\gamma_1,\ \gamma_2)$ is a solution of system (14).

Proof
First, we derive from lemma 5 that if (x_1, x_2) tends from inside S to a boundary point on axis $x_{3-i} = 0$, and x_i^* is a maximum point of $R(x)$ at axis $x_{3-i} = 0$, then

$$\frac{\partial R(x)}{\partial x_{3-i}}\Big|_{x_i = x_i^*} \to \ln \sum_k \frac{\rho_{k,3-i}\alpha_k}{\rho_{k,i}\gamma_i^* + 1}, \tag{43}$$

where γ_i^* is the unique positive solution of equation (17).

Indeed, if $x_{3-i} \to 0$, then from the constraints in lemma 5 and (42) we have:

$$y_k^0 \to \frac{\alpha_k}{1 + \rho_{ki}\gamma_i^0},\ k = 1, 2. \tag{44}$$

Substituting (44) in (40) we get

$$\frac{\partial R(x)}{\partial x_{3-i}} \to \ln \sum_k \frac{\rho_{k,3-i}\alpha_k}{\rho_{k,i}\gamma_i^0 + 1}. \tag{45}$$

If $x_i = x_i^*$, then similarly

$$\frac{\partial R(x)}{\partial x_i}\Big|_{x_i = x_i^*} \to \ln \sum_k \frac{\rho_{k,i}\alpha_k}{\rho_{k,i}\gamma_i^0 + 1} = 0. \tag{46}$$

Now (43) is implied by (45) and (46).

Next, let $i = 1$ and a maximum of $R(x)$ be on axis $x_2 = 0$. By (43) this implies

$$\sum_k \frac{\rho_{k2}\alpha_k}{\rho_{k1}\gamma_1^* + 1} < 1.$$

Then we show that system (14) has a solution with $\gamma_2 < 0$. The case $i = 2$ is similarly proved.

For $t \geq 1$ define a matrix $\|\rho_{ki}(t)\|$ which is obtained from $\|\rho_{ki}\|$ by multiplying its second column by t. Let $(\beta_1(t), \beta_2(t))$ be a solution of the system

$$\sum_k \rho_{ki}(t)\beta_k = 1,\ i = 1, 2$$

and $(\gamma_1(t), \gamma_2(t))$ be a solution of the system

$$\sum_i \rho_{ki}(t)\gamma_i = (\alpha_k - \beta_k(t))/\beta_k(t),\ k = 1, 2.$$

Let $\|\sigma_{ki}\|$ be the inverse matrix for $\|\rho_{ki}\|$. Then $\beta_k(t) = \sigma_{1k} + \sigma_{2k}/t$ and

$$t\gamma_2(t) = \sum_k \sigma_{2k}\left(\frac{\alpha_k}{\beta_k(t)} - 1\right) = \sum_k \left(\sigma_{2k}\frac{\alpha_k}{\sigma_{1k} + \sigma_{2k}/t} - \sigma_{2k}\right).$$

Define

$$t_0 = \left[\sum_k \frac{\rho_{k2}\alpha_k}{\rho_{k1}\gamma_1^* + 1}\right]^{-1}.$$

Note that $0 < \beta_k(t) < \alpha_k$ for all $t \in [1, t_0]$ since $\beta_k(t)$ are monotone functions on $[1, t_0]$ while $\beta_k(1)$ and $\beta_k(t_0) = \alpha_k/(\rho_{k1}\gamma_1^* + 1)$ are in the interval $(0, \alpha_k)$. Define $g(t) = t\gamma_2(t)$. We have: $g(1) = \gamma_2$,

$$g(t_0) = \sum_k \sigma_{2k}\left(\frac{\alpha_k}{\beta_k(t_0)} - 1\right) = \frac{\rho_{11}\rho_{21} - \rho_{21}\rho_{11}}{\Delta}\gamma_1^* = 0$$

and

$$g'(t) = \sum_k \frac{\alpha_k \sigma_{2k}^2}{t^2(\sigma_{1k} + \sigma_{2k}/t)^2}$$

is positive for $t \in [1, t_0]$. Therefore $\gamma_2 < 0$.

5.2 Proof of theorem 1

First, we prove that the function $F(\mathbf{x})$ has a maximum only on the set $V = \{\mathbf{x} \in \mathcal{C} : \sum_{k,i} x_{ki} = v^*\}$, where v^* is the unique root of equation (8) in the interval $(0, \alpha_1 + \alpha_2)$. If $\rho_{k1} = \rho_{k2}$, then (40) implies that

$$\frac{\partial R(x)}{\partial x_1} = \frac{\partial R(x)}{\partial x_2}.$$

Therefore $R(x) = f(x_1 + x_2)$ where $f(v)$ is a smooth function of one variable on $[0, \alpha_1 + \alpha_2]$. By condition $\rho_1\alpha_1 + \rho_2\alpha_2 > 1$ the function $f(v)$ cannot have a maximum at $v = 0$ as $f'(0)) > 0$ by (40). The function $f(v)$ cannot have also a maximum at $v = \alpha_1 + \alpha_2$ because $\partial F(\mathbf{x})/\partial x_{ki} \to -\infty$ as $x_1 + x_2 \to \alpha_1 + \alpha_2$. Therefore $f(v)$ has a maximum at a point $v^* \in (0, \alpha_1 + \alpha_2)$. Consider a set of $x_{ki}^* > 0,\ k, i = 1, 2$ whose sum is v^*. By lemma 4, $x_{ki}^* = \rho_{ki}\gamma_i\beta_k$, (β_1, β_2) satisfy (13) and (γ_1, γ_2) satisfy (14). Under condition $\rho_{ki} \equiv \rho_k$ we have from (14)

$$\beta_k = \frac{\alpha_k}{1 + \rho_k(\gamma_1 + \gamma_2)}, \quad k = 1, 2. \tag{47}$$

By substituting (47) in the first equation of (13) and using the equation $x_i^* = \gamma_i,\ i = 1, 2$, we see that v^* satisfies equation (8). To complete the proof note that equation (8) has a single solution in the interval $(0, \alpha_1 + \alpha_2)$ as under conditions of the theorem the left hand side of (8) denoted by $h(v)$ has the following properties: $h(0) > 1$, $h(\alpha_1 + \alpha_2) < 1$ and $h'(v) < 0$ for $v \in (0, \alpha_1 + \alpha_2)$.

Next, we prove (7), (9) and (10). Assuming $r_{11} = r_{12} = r_1$, $r_{21} = r_{22} = r_2$ and using (1) we have

$$\begin{aligned} P\{Q_{11} = n_{11},\ Q_{12} &= m - n_{11},\ \sum_{k,i} Q_{ki} = l\} \\ &= \frac{1}{G}\frac{N_1!}{(N_1-m)!}\frac{N_2!}{(N_2-(l-m))!} r_2^l \sum_{n_1+n_2=l} \frac{n_1!}{n_{11}!n_{21}!}\frac{n_2!}{n_{12}!n_{22}!}\rho^m \\ &= \frac{r_2^l}{G}\frac{N_1!}{(N_1-m)!}\frac{N_2!}{(N_2-(l-m))!}\binom{l+1}{m+1}\rho^m, \qquad (48) \end{aligned}$$

where $\rho = r_1/r_2$. The last equality is obtained from the identity

$$\binom{l+1}{m+1} = \sum_{i=0}^{l-m}\binom{i+j}{i}\binom{l-j-i}{l-m-i}, \quad 0 \le j \le m \le l. \qquad (49)$$

Using the relation that $\binom{l+1}{m+1} = \binom{l+1}{l-m}$ and rewriting $l-m$ as m, identity (49) can be rewritten as

$$\binom{l+1}{m} = \sum_{i=0}^{m}\binom{i+j}{i}\binom{l-j-i}{m-i}, \quad 0 \le m+j \le l,\ 0 \le m, j.$$

Further replacing l by $l+j+m$, we have

$$\binom{l+j+m+1}{m} = \sum_{i=0}^{m}\binom{i+j}{i}\binom{l+m-i}{m-i}, \quad m, l, j \ge 0.$$

This formula is nothing but (12.16) in [3, page 65]. Apparently, the formula represents the fact that the convolution of negative binomial distributions is binomial.

¿From (48) we have

$$P\{Q_{11}+Q_{12} = m,\ \sum_{k,i} Q_{ki} = l\} = (m+1)\frac{r_2^l}{G}\frac{N_1!}{(N_1-m)!}\frac{N_2!}{(N_2-(l-m))!}\binom{l+1}{m+1}\rho^m \qquad (50)$$

and

$$P\{Q_{11} = j | Q_{11}+Q_{12} = m,\ \sum_{k,i} Q_{ki} = l\} = \frac{1}{m+1},\ j = 0, 1, ..., m. \qquad (51)$$

Similar to [11], one can derive from (50) using assumptions (3) the following asymptotic representation:

$$p(l,m) = P\{Q_{11}+Q_{12} = m,\ \sum_{k,i} Q_{ki} = l\} = c(N)\exp\{N\Psi(v,u) + O(\ln N)\}, \qquad (52)$$

where $v = l/N,\ u = m/N$), $c(N)$ does not depend on (u, v) and

$$\begin{aligned}\Psi(v,u) &= v\ln\rho_2 - v + v\ln v + u\ln\rho - (\alpha_1 - u)\ln(\alpha_1 - u) - u\ln u \\ &- (v^* - u)\ln(v^* - u) - (\alpha_2 - v^* + u)\ln(\alpha_2 - v^* + u). \qquad (53)\end{aligned}$$

We prove that the function $\Psi(v,u)$ has a unique maximum inside $\Gamma = \{(v,u) : v \in (0, \alpha_1 + \alpha_2),\ u \in (0, \alpha_1),\ (v - u) \in (0, \alpha_2)\}$, and the maximum point (v^*, u^*) is defined by the unique solution of equation (8) and (10). Indeed, the function $\Phi(v,u) = -\Psi(v,u)$ is convex in Γ because its second derivative

$$\Phi_{vv} = \frac{1}{v-u} = \frac{1}{\alpha_2 - (v-u)} - \frac{1}{v} > 0,\ (v,u) \in \Gamma$$

and the determinant of the second derivatives

$$||\Phi_{vu}|| = \left(\frac{1}{u} - \frac{1}{v}\right)\left(\frac{1}{v-u} - \frac{1}{\alpha_2 - (v-u)}\right) > 0,\ (v,u) \in \Gamma.$$

Hence, the function $\Psi(v,u)$ has a single maximum inside Γ if the system of two equations, defined by the necessary conditions for a local maximizer, has a solution in Γ. The necessary conditions for a local maximizer of $\Psi(v,u)$ have the following form:

$$\begin{aligned}\Psi'_v(v,u) &= \ln\rho_2 + lnv + - \ln(v-u) + \ln(\alpha_2 - (v-u)) = 0 \\ \Psi'_u(v,u) &= \ln\rho - lnu + \ln(\alpha_1 - u) + \ln(v-u) - \ln(\alpha_2 - (v-u)) = 0\end{aligned}$$

or

$$\frac{v-u}{\alpha_2 - (v-u)} = \rho_2 v, \qquad (54)$$

$$\frac{\alpha_1 - u}{u}\cdot\frac{v-u}{\alpha_2 - (v-u)} = \frac{\rho_2}{\rho_1}. \qquad (55)$$

Substitution of $\rho_2 v$ instead of the second fraction in the left hand side of (55) yields $\alpha_1/u - 1 = 1/(\rho_1 v)$ or

$$u = \frac{\alpha_1\rho_1 v}{1 + \rho_1 v} = \alpha_1 - \frac{\alpha_1}{\rho_1 v + 1} \qquad (56)$$

Consider equation

$$\frac{(\alpha_1 - u)(v^* - u)}{u(\alpha_2 - v^* + u)} = \frac{\rho_2}{\rho_1}. \qquad (57)$$

which is obtained from (55) by the substitution $v = v^*$. By (56) the system of equations (54) and (55) has a solution $(v^*, u^*) \in \Gamma$ if equation (57) has a solution in the interval $(0, \alpha_1)$ given by (10). To prove that (10) satisfies (57) we use the following relations:

$$v^* = \frac{\alpha_1 - \beta_1}{\beta_1 \rho_1} = \frac{\alpha_2 - \beta_2}{\beta_2 \rho_2} \tag{58}$$

and

$$u^* = \alpha_1 - \beta_1. \tag{59}$$

Using (58) and (59) we have

$$v^* - u^* = \frac{\alpha_1 - \beta_1}{\beta_1 \rho_1} - (\alpha_1 - \beta_1) = (\alpha_2 - \beta_2)\frac{1 - \beta_1 \rho_1}{\beta_2 \rho_2} = \alpha_2 - \beta_2, \tag{60}$$

where the last equality is derived using the relation $\rho_1 \beta_1 + \rho_2 \beta_2 = 1$. Substituting (60) and (59) in the left hand side of (57) and using (58) we get ρ_2/ρ_1 that is the right hand side of (57).

Finally, using the representation (52) and the fact that the function $\Psi(v, u)$ has a unique maximum (v^*, u^*) inside Γ, one can prove similar to [11] that

$$\left(\frac{1}{N} \sum_{k,i} Q_{ki},\ \frac{1}{N} \sum_{i} Q_{1i} \right) \xrightarrow{P} (v^*,\ u^*). \tag{61}$$

Convergence in probability (7) and (9) is implied by convergence in distribution because the limit in (61) is deterministic. Now (61) and (51) imply the convergence (11) to the uniform distribution. (12) is similarly proved.

5.3 Proof of theorem 2

We prove below that under conditions of statements 1, 3 and 4 the function $F(\mathbf{x})$ has a unique maximum point $\mathbf{x}^*$ on $\mathcal{C}$ and identify its positive and zero components. Moreover, we prove that $F(\mathbf{x})$ cannot have a maximum at a point, where $x_{ki} = 0$ while $x_i > 0$. Hence representation (4) implies the convergence $\mathbf{Q}/N \xrightarrow{P} \mathbf{x}^*$, [11], while the positive and zero components of $\mathbf{x}^*$ identify the bottleneck and non-bottleneck PS-stations respectively.

1. $F(\mathbf{x})$ is a continuous function on a closed set $\mathcal{C}$, and, therefore, it has a maximum on $\mathcal{C}$. However, $F(\mathbf{x})$ cannot have a maximum on the boundary $x_1 + x_2 = \alpha_1 + \alpha_2$ or at a point, where $x_{ki} = 0$ while $x_i > 0$. The first statement is true because $\partial F(\mathbf{x})/\partial x_{ki} \to -\infty$ as $x_1 + x_2 \to \alpha_1 + \alpha_2$. The second statement is true because by lemma 5, under conditions (39) the function G cannot have a maximum in $D(x)$ when one of $x_{ki} = 0$. Moreover, by lemma 6 $F(\mathbf{x})$ cannot have a maximum at the boundaries $x_1 = 0$ or $x_2 = 0$. Thus $F(\mathbf{x})$ has a maximum point $\mathbf{x}^*$ inside $\mathcal{C}$ with $x^*_{ki} > 0$. By lemma 4 the maximum point is unique and given by (15) as $\Delta \neq 0$.

2. If equations (13) have a solution $\beta_k \in (0, \alpha_k)$, then the condition $\Delta = 0$ implies

$\rho_{ki} \equiv \rho_k$, and the results of the statement follow from theorem 1.
3. We consider cases **A** and **B** separately.
A. Here we prove that the function $F(\mathbf{x})$ has in $\mathcal{C}$ the unique maximum $(x^*_{1i}, x^*_{2i}, 0, 0)$ defined by equations (18) and (17).

Since $\sum_k \rho_{k2}\alpha_k \leq 1$ the system of equation (13) does not have a positive solution with $\beta_k < \alpha_k$. By lemma 4 $F(\mathbf{x})$ cannot have a maximizing $\mathbf{x}^*$, where x^*_{ki} are all positive. $F(\mathbf{x})$ is a continuous function on a closed set $\mathcal{C}$ and, therefore, it has a maximum on $\mathcal{C}$. However, by the same arguments as before $F(\mathbf{x})$ cannot have a maximum on the boundary $x_1 + x_2 = \alpha_1 + \alpha_2$ or at a point, where $x_{ki} = 0$ while $x_i > 0$. Thus $F(\mathbf{x})$ has a maximum at the boundaries $x_1 = 0$ or $x_2 = 0$.

Similar to the proof of lemma 4 we consider the following auxiliary problem: Maximize

$$H_i(x_i, y, x_{1i}, x_{2i}) = (x_i \ln x_i - x_i) + \sum_k x_{ki} \ln \rho_{ki} - \sum_k x_{ki} \ln x_{ki} - \sum_k y_k \ln y_k \quad (62)$$

subject to constraints

$$\sum_k x_{ki} = x_i, \;\; x_{ki} + y_k = \alpha_k, \;\; k = 1, 2$$

on the set

$$x_i > 0, \;\; y_1 > 0, \;\; y_2 > 0.$$

The necessary conditions for a local maximizer for the auxiliary problem give

$$\gamma_i = x^*_i, \;\; \beta_k = y^*_k, \;\; x^*_{ki} = \rho_{ki}\gamma_i\beta_k, \;\; k = 1, 2. \quad (63)$$

We have from (63) and the definition of x_i that (β_1, β_2) satisfy the system (13). Furthermore, (63) and condition $x^*_{ki} + y^*_k = \alpha_k$ imply that

$$\beta_k = \frac{\alpha_k}{\rho_{ki}\gamma_i + 1}, \;\; k = 1, 2. \quad (64)$$

Substituting (64) in (13) we obtain equation (17), where the lower index in γ is omitted. Denote the left hand side of equation (17) by $\phi_i(\gamma)$. Function $\phi_i(\gamma)$ in monotonically decreasing on $[0, \alpha_1 + \alpha_2]$ and $\phi_i(0) = \sum_k \rho_{ki}\alpha_k$. Therefore, if $\sum_k \rho_{ki}\alpha_k < 1$, then equation (17) does not have a solution. This means that the function $F(\mathbf{x})$ may have a maximum only when $x_i = 0$. If $\sum_k \rho_{ki}\alpha_k = 1$, then equation (17) has the only solution at $x_i = 0$. This implies that the function $F(\mathbf{x})$ may have a maximum only when $x_i = 0$. If $\sum_k \rho_{ki}\alpha_k > 1$, then equation (17) has the unique solution γ^*_i since $\phi(\alpha_1 + \alpha_2) < 1$. Substituting $\gamma_i = \gamma^*_i$ in (64) we obtain β^*_k and finally (see (63)) positive components of the maximum point $x^*_{ki} = \gamma^*_i\beta^*_k$ that gives (18).
B. Here we prove that the function $F(\mathbf{x})$ has the unique maximum at axis $x_{3-i} = 0$ if condition (16) is satisfied. (18) is proved similarly to that in **A**.

Since the conditions of statements 1 and 2 are not satisfied, maximum of the function $F(\mathbf{x})$ is at one of the axes $x_1 = 0$ or $x_2 = 0$. Condition $\sum_k \rho_{ki}\alpha_k > 1$, $k = 1, 2$ implies that the function $F(\mathbf{x})$ cannot have a maximum at the origin. This is because of relation (38), where both partial derivatives of the function $R(x)$ given by (40) are positive at $x = 0$. Moreover, one can prove as before that $F(\mathbf{x})$ cannot have a maximum on the boundary $x_1 + x_2 = \alpha_1 + \alpha_2$ or at a point, where $x_{ki} = 0$ while $x_i > 0$.

First, we consider the case when system (13) does not have a solution $\beta_k \in (0, \alpha_k)$. Define the two following sets

$$Z = \{\mathbf{z} : 0 \leq z_k \leq \alpha_k,\ \sum_k \rho_{k1} z_k = 1\},$$

$$W = \{\mathbf{w} : 0 \leq w_k \leq \alpha_k,\ \sum_k \rho_{k2} w_k = 1\}.$$

Our assumption implies that either

$$\max_{\mathbf{z}\in Z} \sum_k \rho_{k2} z_k < 1 \tag{65}$$

or

$$\min_{\mathbf{z}\in Z} \sum_k \rho_{k2} z_k > 1. \tag{66}$$

It is easy to see that for $i = 1$, condition (16) is satisfied if and only if (65) holds. For $i = 2$, condition (16) is satisfied if and only if (66) holds. This is because our assumption implies that $\max_{\mathbf{w}\in W} \sum_k \rho_{k1} w_k < 1$ if (66) holds, and $\min_{\mathbf{w}\in W} \sum_k \rho_{k1} w_k > 1$ if (65) holds. Hence under condition (16), the maximum of function $R(x)$ can be only on axis $x_{3-i} = 0$ because this is the only case when the partial derivative (43) is negative.

Next, we consider the case when $\Delta \neq 0$, and system (13) has a solution $\beta_k \in (0, \alpha_k)$ but system (14) has a solution with $\gamma_1 \leq 0$ or $\gamma_2 \leq 0$. Hence by lemmas 4 and 5 the function $R(x)$ does not have a maximum inside S. By lemma 6 the function $R(x)$ cannot have local maximums on both axes simultaneously. These two facts prove the statement.

4. It was proved in **A** that if $\sum_k \rho_{ki}\alpha_k < 1$, then function $F(\mathbf{x})$ may have a maximum only when $x_i = 0$. This implies that $\mathbf{x}^* = \mathbf{0}$ is the unique maximum of $F(\mathbf{x})$ on the set $\mathcal{C}$.

6 Generalizations

In section 3 we covered the bottleneck analysis for all cases that can occur when $K = M = 2$. When $M > 2$ and $K \geq 2$ the bottleneck analysis becomes more complicated as many more cases are possible and their complete exposition is beyond the scope of this paper. However, in this section we state the results for an important subset of the cases, where bottleneck groups are excluded. Generalizations described in this section are based on comparison of the case of $K = M = 2$ with the previously studied cases of $K = 1,\ M > 1$ and $M = 1,\ K > 1$ [7, 11, 9]. These generalizations can be proved by the same technique that is used in section 5 for $K = M = 2$.

In the case $K = 1,\ M > 1$, the network consists of one IS station and M single servers (SS) numbered as $1, \ldots, M$. SS 1 is a bottleneck if (see [7])

$$\rho_1 > 1 \quad \text{and} \quad \rho_1 > \rho_i,\ i = 2, \ldots, M. \tag{67}$$

(Only the second index is used here since $k = 1$.) When there are multiple classes, the load at PS station i is defined as a linear combination of the per-class loads, and station i could be a bottleneck or belong to a bottleneck group if its load exceeds 1:

$$\sum_k \rho_{ki}\alpha_k > 1. \tag{68}$$

In the case $M = 1,\ K > 1$, the network consists of one IS station with K classes and one PS station. The bottleneck condition (see [11, 9]) is $\sum_k \rho_k\alpha_k > 1$. (Only the first index is used here since $i = 1$.) Under this condition equation (cf. (17))

$$\sum_k \frac{\rho_k\alpha_k}{\rho_k\gamma + 1} = 1 \tag{69}$$

has the unique positive solution γ^*, and x_k^* is given by equation (18) with omitted index i.

When both $K, M > 1$ the bottleneck identification becomes more complicated as one can see from the previous section. This is because more than one bottleneck may exist and, in general, ordering of the loads for all PS stations (cf. (67)) does not identify the bottleneck even in the case of one bottleneck (see **B** in statement 3 of theorem 2) Note that **A** in statement 3 and statement 4 of theorem 2) apply for arbitrary K, M and are already so stated. Moreover, statement 1 in theorem 2 directly generalizes for $K > 2$ if $K = M$.

In general, one can solve (14) with respect to β_k and substitute this solution in (13), which gives one system of nonlinear equations

$$\sum_k \frac{\rho_{ki}\alpha_k}{1 + \sum_j \rho_{kj}\gamma_j} = 1 \tag{70}$$

with respect to γ_i instead of two systems of linear equations. Equations (70) are valid only for those i that correspond to bottleneck PS stations for which $\gamma_i > 0$. For non-bottleneck nodes $\gamma_j = 0$ in (70). To identify the bottleneck nodes it is necessary to find the maximum subset $B_L = (i_1, \ldots, i_L)$ of L ($\leq K$) PS stations for which the condition (68) is satisfied, and the system

$$\sum_k \frac{\rho_{ki}\alpha_k}{1 + \sum_{j \in B_L} \rho_{kj}\gamma_j} = 1, \quad i \in B_L \tag{71}$$

has a positive solution $\{\gamma_i^* > 0,\ i \in B_L\}$ such that

$$\sum_k \frac{\rho_{ki}\alpha_k}{1 + \sum_{j \in B_L} \rho_{kj}\gamma_j^*} < 1, \quad i \notin B_L \tag{72}$$

(cf. (16) and (17)). These two conditions can be used for bottleneck identification for any M and K, and, in particular, in the case when one of the systems of linear equations in the generalization of statement 1 in theorem 2 does not have a positive solution, i.e. not all PS stations are bottlenecks. Moreover, for $K > M$ they cannot be simplified.

The maximum number of bottleneck stations and groups is $\min(M, K)$. When $M > K$ one can find K bottleneck nodes (if they exist) by solving linear equations (13) and (14) for different subsets $B_K = (i_1, \ldots, i_K)$ of K nodes, i.e. $i \in B_K$ in (13) and (14). The subset B_K^* is the bottleneck subset if both systems have positive solutions for $i \in B_K^*$ and, in addition,

$$\sum_k \rho_{ki}\beta_k^* < 1, \quad i \notin B_K^*. \tag{73}$$

For a single class network $K = 1$ and under conditions (67), $B_1 = (1)$ and $\beta_1 = 1/\rho_1$ which in turn implies inequalities (73). However, we do not know whether condition (73) can be further simplified for a general multiple class network.

Finally, for $M > 2$ more than one bottleneck group may exist and theorem 1 can be generalized for each group. Thus in general, a multiple class network with multiple PS stations may have bottleneck nodes, bottleneck groups and non-bottleneck nodes.

7 Acknowledgment

We would like to thank Ward Whitt for his insightful discussions on the results of the paper, and Kathy Meier-Hellstern, Dave Houck, and Pat Wirth for their careful review of an earlier draft of the paper. We also thank the referee for helpful comments and probabilistic interpretation of identity (49) which we included in the paper.

References

[1] A. Berger and Y. Kogan "Dimensioning bandwidth for elastic traffic high-speed data networks", *Submitted for publication.*

[2] A. Birman and Y. Kogan, "Asymptotic evaluation of closed queueing networks with many stations", *Commun.Statist.-Stochastic Models* **8**, 543-564, 1992.

[3] W. Feller, *An Introduction to Probability Theory and Its Application*, vol. 1, Third edition, Revised printing, J. Wiley & Sons, New York, 1970.

[4] M. Freidlin and A. Wentzell, *Random Perturbations of Dynamical Systems*, Springer-Verlag, New York, 1984.

[5] F.P. Kelly, "The dependence of sojourn times in closed queueing networks", in G. Iazeollla, P.J. Courtois and A. Hordijk (eds), *"Mathematical Computer Performance and Reliability,* 111-121, North-Holland, Amsterdam, 1984.

[6] Y. Kogan, "Another approach to asymptotic expansions for large closed queueing networks", *Oper.Res.Letters* **11**, 317-321, 1992.

[7] Y. Kogan and A. Birman, "Asymptotic analysis of closed queueing networks with bottlenecks", in T.Hasegawa, H.Takagi and Y. Takahashi (eds.), *IFIP Transactions C-5, Performance of Distributed Systems and Integrated Communication Networks*, 265-280, North-Holland, Amsterdam, 1992.

[8] Y. Kogan and A. Yakovlev, "Asymptotic analysis for closed multichain queueing networks with bottlenecks", *Queueing Systems* **23**, 235-258, 1996.

[9] J. McKenna and D. Mitra, "Integral representation and asymptotic expansions for closed Markovian queueing networks: normal usage" *Bell Syst. Tech. J.* **61**, 661-683, 1982.

[10] A. L. Peressini, F. E. Sullivan, J. J. Uhl, Jr. *The Mathematics of Nonlinear Programming*, Springer-Verlag, New York, 1988.

[11] B. Pittel, "Closed exponential networks of queues with saturation: The Jackson-type stationary distribution and its asymptotic analysis", *Math.Oper.Res.* **6**, 357-378, 1979.

[12] Private Network–Network Interface Specification, Version 1.0, ATM Forum, March 1996.