New Linear Program Performance Bounds for Queueing Networks*

James R. Morrison† and P. R. Kumar‡

Abstract

We obtain new linear programs for bounding the performance and proving the stability of queueing networks. They exploit the key facts that the transition probabilities of queueing networks are shift invariant on the relative interiors of faces and the cost functions of interest are linear in the state. A systematic procedure for choosing different quadratic functions on the relative interiors of faces to serve as surrogates of the differential costs in an inequality relaxation of the average cost function leads to linear program bounds. These bounds are provably better than earlier known bounds. It is also shown how to incorporate additional features such as the presence of virtual multiserver stations to further improve the bounds. The approach also extends to provide functional bounds valid for all arrival rates.
Keywords: Queueing networks, scheduling, stability, performance evaluation.

1 Introduction

In [1, 2] linear programs are provided to bound the performance of queueing networks.

The duality of the LP in [1] to a stability establishing LP in [3] was shown in [4].

*This material is based upon work partially supported by the Semiconductor Research Corporation under Contract No. 97-FJ-489, the U.S. Army Research Office under Contract No. DAAH-04-95-1-0090, the National Science Foundation under Grant Nos. ECS-94-03571 and DMI-9743165, and the Joint Services Electronics Program under Contract No. N00014-96-1-0129. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the above agencies. The authors are grateful to Lars-Erik Andersson for useful comments related to the nonnegativity of quadratic expressions in orthants.

†Email:morrison@decision.csl.uiuc.edu

‡Professor, University of Illinois, Coordinated Science Laboratory, 1308 West Main Street, Urbana, IL 61801. Email: prkumargauss.csl.uiuc.edu. Tel: (217) 333-7476.

Functional bound LPs were demonstrated for closed networks in [5] and open networks in [6].

In this paper we obtain new LPs which provide superior performance bounds, upper or lower. They are based on exploiting the key fact that the transition probabilities for queueing networks are shift invariant in the relative interiors of faces. This allows for a systematic construction of separate quadratic functions in each such region which serve as surrogates for the differential cost functions in an inequality relaxation of the average cost equation of dynamic programming [7]. The result is LPs which provide provably better bounds than earlier. Several examples are provided to show their superiority to earlier bounds.

We also show how to incorporate the presence of any virtual multiserver stations, see [8, 9, 10], into the bounding procedure to further improve the bounds.

The procedure also extends to provide functional bounds valid over a range of arrival rates just by solving a single LP.

2 Description of queueing network

To expose the ideas in their simplest form, we consider open re-entrant lines – the approach applies to more general networks as in [1].

An open re-entrant line is a queueing network consisting of S stations $\sigma_1, \sigma_2, \ldots, \sigma_S$, at which customers receive service, and L buffers $b_1, b_2, \ldots, b_L$, at which they await service. Buffer b_i is serviced at station $\sigma(i) \in \{\sigma_1, \ldots, \sigma_S\}$, and we write $i \in \sigma$ to denote that $\sigma(i) = \sigma$. Customers arrive at buffer b_1 as a Poisson process of rate

λ. Upon completing service from buffer b_i at station $\sigma(i)$, a customer moves next to buffer b_{i+1} at station $\sigma(i+1)$, unless $i = L$ in which case the customer exits the system. The service time of a customer in buffer b_i is exponentially distributed with mean $\frac{1}{\mu_i}$. All service times and interarrival times are independent. The state of the system is $x(t) = (x_1(t), x_2, (t), \ldots, x_L(t))^T$ where $x_i(t)$ is the number of customers in buffer b_i at time t. Let $u_i(t) = 1$ if a customer in buffer b_i is undergoing service at time t, and $u_i(t) = 0$ otherwise. Each station can only serve one customer at any given time. A scheduling policy is called *non-idling* if a station is busy whenever any of its buffers is nonempty. We restrict attention to such non-idling policies. A special case of a non-idling policy is a *buffer priority policy* θ which is given by a priority ordering $(\theta(1), \theta(2), \ldots, \theta(L))$, a permutation of $\{1, 2, \ldots, L\}$, with the understanding that priority is given at a station σ to that nonempty buffer b_i for which $\theta(i) < \theta(j)$ for all other nonempty buffers b_j located at σ. The priority is implemented in a preempt resume manner.

We normalize time so that $\lambda + \sum_{i=1}^{L} \mu_i = 1$, and sample the system at all random times τ_n corresponding to arrivals, and real or virtual service completions, where we assume that each buffer not being served has a virtual customer in service. Let $u_i(t) = 1$ if a customer in buffer b_i is being served by station $\sigma(i)$ at time t, and 0 otherwise. Assuming that each $u_i(t)$ is constant in $[\tau_n, \tau_{n+1})$, we obtain a controlled Markov chain whose state, denoted $x(n)$, is the vector $x(\tau_n)$ of the continuous time process. Similarly, we denote by $u(n)$ the control vector $u(\tau_n)$. If $e_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0)^T$ is the unit vector with a 1 in the ith position, and e_{L+1} is the zero vector, then the controlled

transition probabilities are:

$$\begin{aligned} p_{x,x+e_1}(u) &= \lambda \\ p_{x,x-e_i+e_{i+1}}(u) &= \mu_i u_i \qquad \text{for } i = 1, 2, \ldots, L. \end{aligned}$$

Note that $u_i = 1$ is feasible above only if $x_i \geq 1$. Above $x \in Z_+^L$.

3 The average cost inequality

We start with a well known idea, consisting of relaxing the average cost equation to an inequality.

Lemma 3.1. *Consider a stationary control policy $u(x)$. Let $c(x)$ be a given one-step cost function. Let $x(0)$ be an arbitrary initial condition.*

(i) *Suppose there is a real number J, and a function $W(x)$ which is bounded below such that*

$$J + W(x) \geq c(x) + \sum_{y \in X} p_{x,y}(u(x)) W(y) \text{ for all } x. \tag{1}$$

Then

$$J \geq \limsup_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} Ec(x(n)). \tag{2}$$

(ii) *Suppose there is a real number J, and a function $W(x)$ which is bounded above such that*

$$J + W(x) \leq c(x) + \sum_{y \in X} p_{x,y}(u(x)) W(y) \text{ for all } x. \tag{3}$$

Then

$$J \quad \leq \quad \liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} Ec(x(n)). \tag{4}$$

Proof: The inequality in (1) can be written as

$$J + W(x(n)) \quad \geq \quad c(x(n)) + E[W(x(n+1)) \mid x(n)],$$

and so

$$J \quad \geq \quad \frac{1}{N} \sum_{n=0}^{N-1} c(x(n)) + \frac{EW(x(N)) - W(x(0))}{N}.$$

Since $W(x(N))$ is bounded below, taking the limit yields the result. The proof of (ii) is similar. □

For a general system an inequality such as (1) or (3) is formidable to verify since it needs to be satisfied for all x. However for queueing networks a key property is that the transition probabilities are shift invariant in the relative interiors of faces. That is, for any $y \geq x$ (componentwise) in the relative interior of the face, and any u feasible for x, which is therefore also feasible for y,

$$p_{x,x+e_1}(u) \quad = \quad p_{y,y+e_1}(u) = \lambda$$

$$p_{x,x-e_i+e_{i+1}}(u) \quad = \quad p_{y,y-e_i+e_{i+1}}(u) = \mu_i u_i.$$

A second key fact is that our interest centers on the choice of the cost function $c(x) = |x| := \sum_{i=1}^{L} x_i$, a linear function of x. As a consequence, it turns out that if $W(x)$ is a quadratic function, then verification of an inequality such as (1) or (3) on the infinite number of states in the entire relative interior of a face can be done by merely verifying a finite number of inequalities.

It is shown in the Appendix that the choice $W(x) = \frac{1}{2}x^T Q x + p^T x$ improves upon the linear programs in [1]. To obtain linear programs yielding greater improvement we employ a different quadratic function on the relative interior of each face.

We need to define what we mean by the relative interiors of faces. Let $\phi = (\phi_1, \phi_2, \ldots, \phi_L)^T$ be a binary vector with each $\phi_i = 0$ or 1. Each ϕ labels a region

$$X^\phi \quad := \quad \{x \in \mathcal{Z}_+^L \ : \ x_i \geq 1 \text{ if } \phi_i = 1, \text{ and } x_i = 0 \text{ if } \phi_i = 0\}$$

of the nonnegative orthant of the integer lattice where only certain components are nonzero. In each such region we allow a separate quadratic

$$W^\phi(x) \quad = \quad \frac{1}{2}x^T Q^\phi x + p^{\phi T} x + c^\phi, \quad (Q^\phi = \text{symmetric}), \tag{5}$$

and consider the composite

$$W(x) \quad = \quad W^\phi(x) \text{ if } x \in X^\phi.$$

As will be shown, only a finite number of inequalities need to be checked to guarantee (1) or (3).

4 Buffer priority policies

Buffer priority policies are easier to consider than other stationary policies since in each X^ϕ only one choice of u is made; call it $u(\phi)$. For each x, let $\phi(x)$ denote that label for which $x \in X^{\phi(x)}$. To verify (1) we need to ensure that

$$\begin{aligned} J - |x| \quad \geq \quad & \lambda[c^{\phi(x+e_1)} + p^{\phi(x+e_1)T}(x + e_1) + \frac{1}{2}(x + e_1)^T Q^{\phi(x+e_1)}(x + e_1) \\ & -c^{\phi(x)} - p^{\phi(x)T} x - \frac{1}{2}x^T Q^{\phi(x)} x] \end{aligned}$$

$$+\sum_{i:u_i(\phi(x))=1}\mu_i[c^{\phi(x-e_i+e_{i+1})}+p^{\phi(x-e_i+e_{i+1})^T}(x-e_i+e_{i+1})$$
$$+\frac{1}{2}(x-e_i+e_{i+1})^TQ^{\phi(x-e_i+e_{i+1})}(x-e_i+e_{i+1})-c^{\phi(x)}-p^{\phi(x)^T}x$$
$$-\frac{1}{2}x^TQ^{\phi(x)}x] \text{ for all } x. \qquad (6)$$

We want to simplify this by exploiting the key fact that $\phi(x)$ does not vary with x unless the set of nonzero components is changed. If $x_1 = 0$, then $\phi(x+e_1)$ is different from $\phi(x)$. Similarly if $x_i = 1$ or $x_{i+1} = 0$, then $\phi(x-e_i+e_{i+1})$ differs from $\phi(x)$. Thus, a finer subdivision of X^ϕ is called for.

Let $\psi = (\psi_1, \psi_2, \ldots, \psi_L)^T$ be a vector with each $\psi_i \in \{0,1,2\}$. Each such ψ labels a region

$$Y^\psi \quad = \quad \{x \in Z_+^L : x_i = 0 \text{ if } \phi_i = 0, x_i = 1 \text{ if } \psi_i = 1, \text{ and } x_i \geq 2 \text{ if } \psi_i = 2\}.$$

Each Y^ψ is a subset of $X^{\phi(\psi)}$.

Let D^ψ denote the diagonal matrix with $d_{ii}^\psi = 1$ if $\psi_i = 2$, and 0 otherwise. Then for $x \in Y^\psi$, $D^\psi x$ extracts from x only those components which are equal to 2 in ψ. Thus, every vector $x \in Y^\psi$ can be written as $x = \psi + D^\psi z$ for some $z \in \mathcal{Z}_+^L$. Hence $x \in Y^\psi \Leftrightarrow x = \psi + D^\psi z$ for some $z \in Z_+^L$, and moreover for such x,

$$\phi(x+e_1) \quad = \quad \phi(\psi+e_1),$$
$$\phi(x-e_i+e_{i+1}) \quad = \quad \phi(\psi-e_i+e_{i+1}).$$

This renders the above ϕ's dependent only on the label ψ, and not x.

In turn, this allows us to simplify (6). We need only ensure that for each ψ,

$$J-|\psi+D^\psi z| \quad \geq \quad \lambda[c^{\phi(\psi+e_1)}+p^{\phi(\psi+e_1)^T}(\psi+D^\psi z+e_1)$$

$$
\begin{aligned}
&\frac{1}{2}(\psi + D^\psi z + e_1)^T Q^{\phi(\psi+e_1)}(\psi + D^\psi z + e_1) \\
&-c^{\phi(\psi)} - p^{\phi(\psi)T}(\psi + D^\psi z) - \frac{1}{2}(\psi + D^\psi z)^T Q^{\phi(\psi)}(\psi + D^\psi z)] \\
&+ \sum_{i:u_i(\phi(\psi))=1} \mu_i [c^{\phi(\psi - e_i + e_{i+1})} + p^{\phi(\psi - e_i + e_{i+1})T}(\psi + D^\psi z - e_i + e_{i+1}) \\
&+\frac{1}{2}(\psi + D^\psi z - e_i + e_{i+1})^T Q^{\phi(\psi - e_i + e_{i+1})}(\psi + D^\psi z - e_i + e_{i+1}) \\
&-c^{\phi(\psi)} - p^{\phi(\psi)T}(\psi + D^\psi z) - \frac{1}{2}(\psi + D^\psi z)^T Q^{\phi(\psi)}(\psi + D^\psi)] \\
&\text{for all } z \in \mathcal{Z}_+^L.
\end{aligned}
$$

Since the right-hand side above is a quadratic in $D^\psi z$, and the left-hand side is affine in $D^\psi z$, this reduces to checking a quadratic inequality

$$
J \quad \geq \quad d^\psi + r^{\psi T} D^\psi z + \frac{1}{2} z^T D^\psi S^\psi D^\psi z \quad \text{for all } z \in \mathcal{Z}_+^L. \tag{7}
$$

The constant d^ψ, the vector r^ψ, and the symmetric matrix S^ψ are specified in Appendix A.2.

To ensure that (7) holds, we can require that

$$
J \quad \geq \quad d^\psi \tag{8}
$$

$$
r^{\psi T} D^\psi \quad \leq \quad 0 \text{ (componentwise)} \tag{9}
$$

$$
D^\psi S^\psi D^\psi \text{ is conegative.}
$$

Above, a symmetric matrix A is *conegative* if $x^T A x \leq 0$ for all $x \in R_+^L$, i.e., in the nonnegative orthant. The last condition is met by the more stringent condition

$$
D^\psi S^\psi D^\psi \quad \leq \quad 0 \text{ (componentwise).} \tag{10}
$$

The inequalities (8, 9, 10) are linear in the variables c^ϕ, p^ϕ, and Q^ϕ.

We also want to ensure that each $W^\phi(x)$ is bounded below in X^ϕ. An easy though restrictive way to do this is as follows (a less restrictive way will be shown later). We write $x \in X^\phi$ as $x = \phi + D^\phi z$ for some $z \in \mathcal{Z}_+^L$ where D^ϕ is a diagonal matrix with $d_{ii}^\phi = 1$ if $\phi_i = 1$, and 0 otherwise. Then

$$\begin{aligned} W^\phi(x) &= (1/2)z^T D^\phi Q^\phi D^\phi z + \phi^T Q^\phi D^\phi z + (1/2)\phi^T Q^\phi \phi \\ &\quad + p^{\phi T}\phi + p^{\phi T} D^\phi z + c^\phi. \end{aligned}$$

This will be bounded below by the constant terms in the above expression if

$$\overline{p}^\phi := D^\phi p^\phi + D^\phi Q^\phi \phi \geq 0 \text{ (componentwise)}, \tag{11}$$

and

$$\overline{Q}^\phi := D^\phi Q^\phi D^\phi \geq 0 \text{ (componentwise)}. \tag{12}$$

This allows linear programming to determine a performance bound.

Theorem 4.1: (Upper and lower bounds for buffer priority policies). *Fix a buffer priority policy $u(\cdot)$ and consider the one step cost function $c(x) = \mid x \mid$.*

(i) *Let $\{c^\phi, p^\phi$ and symmetric $Q^\phi\}$ be the decision variables in the linear program:*

$$\text{Min} \quad J$$

subject to (8, 9, 10) for each ψ and (11, 12) for each ϕ. Its value J is an upper bound as in (2).

(ii) *The value of the LP*

$$\text{Max} \quad J$$

subject to

$$\begin{aligned} J &\leq d^{\psi} \\ r^{\psi T} D^{\psi} &\geq 0 \ \textit{(componentwise)} \\ D^{\psi} S^{\psi} D^{\psi} &\geq 0 \ \textit{(componentwise)} \end{aligned}$$

for each ψ, is a lower bound as in (4).

Proof: Lemma 3.1 proves (i). For (ii), if the first moment is not finite, then any bound obtained by the linear program is trivially a lower bound. Thus, assume a finite first moment. The proof of Theorem 1 of [4], shows that under this assumption $\lim_{N\to\infty} E \mid x(N) \mid^2 /N = 0$, which is sufficient to show $\lim_{N\to\infty} E[W(x(N))]/N = 0$ (since $W(x)$ is merely a glorified quadratic). Hence, the need for $W(x)$ to be bounded above in the proof of Lemma 3.1 (ii) is eliminated. □

Above we have taken the easiest, and most restrictive path, to guaranteeing that the inequalities of (7) hold, and that $W(x)$ is bounded below. For (7), the basic issue is of ensuring that a quadratic maintains a certain sign throughout the positive orthant. A sharper approach is to use the following characterization.

Lemma 4.2. *A quadratic $c + p^T x + \frac{1}{2} x^T Q x$ is nonnegative over all $x \in R_+^L$ if and only if*

$$c \geq 0 \tag{13}$$

$$Q \text{ is copositive}, \tag{14}$$

and

$$x^T(cQ - pp^T/2)x \geq 0 \quad \text{for all} \quad x \in R^L_+ \text{ with } p^T x < 0. \tag{15}$$

Proof: Since $x = 0$ is a valid choice, (13) is necessary. For large x, the quadratic term dominates the linear and constant terms. Hence (14) is also needed. The restriction on the linear term is a bit more indirect. Clearly, if $p^T x \geq 0$, then (13, 14) are sufficient to ensure nonnegativity at x. Hence consider an $x \in R^L_+$ with $p^T x < 0$. Scaling x by $\alpha \geq 0$ yields $c + \alpha(p^T x) + \alpha^2 \left(\frac{1}{2} x^T Q x\right)$, a quadratic in α. Since Q is copositive and $x \in R^L_+, \frac{1}{2} x^T Q x \geq 0$. Hence the quadratic in α is convex. Requiring the minimum over $\alpha \geq 0$ to be nonnegative yields (15). □

The issue of a quadratic function being bounded below is addressed by the following lemma.

Lemma 4.3. *Consider a quadratic $f(x) = p^T x + \frac{1}{2} x^T Q x$ for $x \in R^L_+$.*

(i) *If $f(x)$ is bounded below, then Q is copositive and $p^T x \geq 0$ for all x such that $x^T Q x = 0$.*

(ii) *If Q is copositive and $p^T x > 0$ for all x such that $x^T Q x = 0$, then $f(x)$ is bounded below.*

Proof: The necessary condition in (i) is clearly true. To prove (ii), argue by contradiction. If $f(x)$ is not bounded below in R^L_+, then there is a sequence $\{x_n\} \in R^L_+$ with $r_n := ||x_n|| \to +\infty$ satisfying $\delta_n := -f(x_n) \to +\infty$. Without loss of generality, suppose

that δ_n is strictly monotone increasing and strictly positive. Consider the projection $v_n := x_n/r_n$ of x_n on the unit circle. Then $f(x_n) = r_n^2(v_n^T Q v_n + p^T v_n/r_n) = -\delta_n \to -\infty$. Since $v_n^T Q v_n \geq 0$ and $p^T v_n/r_n \to 0$, $\liminf_{n\to\infty}(v_n^T Q v_n + p^T v_n/r_n) \geq 0$. Also, since $-\delta_n/r_n^2 \leq 0$, $\limsup_{n\to\infty}(-\delta_n/r_n^2) \leq 0$. Hence, $\lim_{n\to\infty}(v_n^T Q v_n + p^T v_n/r_n) = 0 = \lim_{n\to\infty} -\delta_n/r_n^2$. Now, $p^T v_n < 0$, since $p^T v_n/r_n = -\delta_n/r_n^2 - v_n^T Q v_n$ and the first term on the right is strictly negative while the second term is nonpositive. Consider any subsequence $v_{n_k} \to v^*$. For such a subsequence, $\lim_{k\to\infty} p^T v_{n_k} \leq 0$. Also, $0 = \lim_{k\to\infty}(v_{n_k}^T Q v_{n_k} + p^T v_{n_k}/r_{n_k}) = \lim_{k\to\infty} v_{n_k}^T Q v_{n_k} = v^{*T} Q v^*$ (where the second equality holds since $p^T v_n$ is bounded). But then, $v^{*T} Q v^* = 0$ and $p^T v^* \leq 0$, contradicting the hypothesis. □

Applying this to our context yields the following theorem.

Theorem 4.4.

(i) *If J satisfies for each ψ and ϕ,*

$$J \geq d^\psi,$$

$$D^\psi S^\psi D^\psi \text{ is conegative,} \tag{16}$$

$$z^T \left[(d^\psi - J)D^\psi S^\psi D^\psi - D^\psi r^\psi r^{\psi T} D^\psi/2\right] z \geq 0 \text{ for}$$

$$\text{each } z \in R_+^L \text{ with } r^{\psi T} D^\psi z > 0, \tag{17}$$

$$\overline{Q}^\phi \text{ is copositive,} \tag{18}$$

$$\overline{p}^{\phi T} z > 0 \text{ for all } z \in R_+^L \text{ with } z^T \overline{Q} z = 0, \tag{19}$$

then J is an upper bound as in (2).

(ii) *If J satisfies for each ψ and ϕ*

$$J \leq d^{\psi},$$

$$D^{\psi}S^{\psi}D^{\psi} \text{ is copositive,} \tag{20}$$

$$z^{T}[(d^{\psi} - J)D^{\psi}S^{\psi}D^{\psi} - D^{\psi}r^{\psi}r^{\psi T}D^{\psi}/2]z \geq 0 \text{ for each}$$

$$z \in R_{+}^{L} \text{ with } r^{\psi T}D^{\psi}z < 0, \tag{21}$$

then J is a lower bound as in (4).

Unfortunately, the conditions (16, 17, 18, 19) or (20, 21) are nonlinear. In fact, testing for copositivity or conegativity is NP-Complete (see [11, 12]). Several alternatives are possible which are more restrictive though. One could just require nonnegativity or nonpositivity, as appropriate, of the matrices in (16, 18, 20). This we have already done in Theorem 4.1. Alternately, we could omit the conditions entirely from the linear program, except for sign constraints on the diagonals of the quadratic matrices which arise and which are anyway necessary, and merely check to see whether the solution obtained satisfies the conditions. For the examples presented in the paper we imposed (8, 9, 10) and then checked the boundedness below condition at the end.

5 Bounds for the class of all non-idling policies

If one wants to obtain performance bounds for the class of all non-idling policies, then one only needs to allow for all feasible choices of u in a state x, or equivalently, for

each ψ, in (7). Thus, one simply obtains more constraints. The bounds naturally are looser.

6 Bounds for linear switching curves

The procedure described above extends to certain scheduling policies where the regions of the state space in which a particular buffer receives service are described by linear constraints upon the state variables. An example illustrates the possibilities.

Example 6.1. *Consider the network of Figure 1. Suppose that buffer b_1 receives service in the subset X^{b_1} of X given by*

$$X^{b_1} = \{x \in X : x_1 \geq 1, x_1 \geq m_{13}x_3, x_1 \geq m_{14}x_4\}.$$

Similarly, let

$$X^{b_3} = \{x \in X : x_3 \geq 1, x_1 < m_{13}x_3, x_3 \geq m_{34}x_4\}$$

$$X^{b_4} = \{x \in X : x_4 \geq 1, x_1 < m_{14}x_4, x_3 < m_{34}x_4\}$$

be the regions in which buffers b_3 and b_4 receive service, respectively, where $m_{ij} \geq 0$. Note that in order to guarantee that every possible state in which station σ_1 cannot idle is covered by $\cup_{i=1,3,4} X^{b_i}$ we require $m_{14} = m_{13}m_{34}$.

As before, propose a separate quadratic on each transition invariant region and use $\phi_{i,j}$ to denote the region in which buffer b_i receives service at station σ_1 and buffer b_j receives service at station σ_2. We use $i = 0$ and $j = 0$ to denote that no buffer is in service at the respective station. Thus, we propose quadratics on the regions

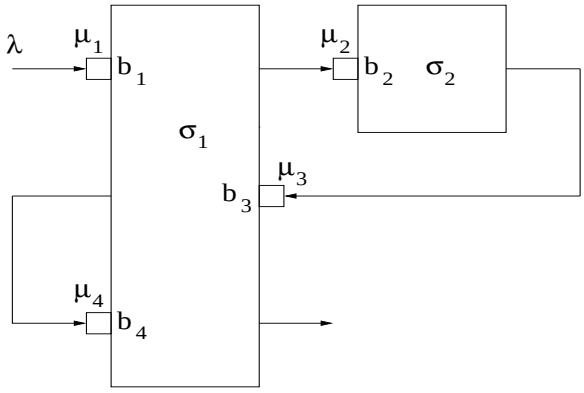

Figure 1: System of Example 6.1

$\phi_{00}, \phi_{10}, \phi_{30}, \phi_{40}, \phi_{02}, \phi_{12}, \phi_{32}$, and ϕ_{42}. *Each region may need to be partioned further to consider the inequalities of Lemma 3.1. Let us examine the region* $\phi_{40} = \{x \in X : x_4 \geq 1, x_2 = 0, x_1 < m_{13}m_{34}x_4, x_3 < m_{34}x_4\}$. *Consider* $x \in \phi_{40}$ *such that* $x_1 + 1 < m_{13}m_{34}x_4$, $x_1 < m_{13}m_{34}(x_4 - 1)$, *and* $x_3 < m_{34}(x_4 - 1)$. *For all such* x, *the next possible states remain in* ϕ_{40}, *so that for every such state we require,*

$$\begin{aligned} J \geq x_1 + x_3 + x_4 \quad &+ \quad \lambda[W^{\phi_{40}}(x + e_1) - W^{\phi_{40}}(x)] \\ &+ \quad \mu_4[W^{\phi_{40}}(x - e_4) - W^{\phi_{40}}(x)]. \end{aligned} \tag{22}$$

Equivalently (suppressing the superscript ϕ_{40}*),*

$$\begin{aligned} J \geq \lambda[p_1 + \frac{1}{2}q_{11}] \quad &+ \quad \mu_4[-p_4 + \frac{1}{2}q_{44}] + x_1[1 + \lambda q_{11} - \mu_4 q_{14}] \\ &+ \quad x_3[1 + \lambda q_{13} - \mu_4 q_{34}] + x_4[1 + \lambda q_{14} - \mu_4 q_{44}]. \end{aligned}$$

We now invoke the linear constraints on x_1, x_3, *and* x_4. *Clearly, we must have*

$$A := 1 + \lambda q_{14} - \mu_4 q_{44} \leq 0,$$

so that requiring (where $\alpha_1, \alpha_3 \geq 0$ *and* $\alpha_1 + \alpha_3 = 1$*)*

$$\begin{aligned} J \geq \lambda[p_1 + \frac{1}{2}q_{11}] \quad &+ \quad \mu_4[-p_4 + \frac{1}{2}q_{44}] + \alpha_1(\frac{x_1}{m_{14}})A + \alpha_3(\frac{x_3}{m_{34}})A \\ &+ \quad x_1[1 + \lambda q_{11} - \mu_4 q_{14}] + x_3[1 + \lambda q_{13} - \mu_3 q_{34}], \end{aligned}$$

will ensure (22). We see here that choices will arise in the construction of a linear program as one is free to choose the values α_1 *and* α_3*. Once such values have been chosen, we can then require*

$$\begin{aligned} J &\geq \lambda[p_1 + \frac{1}{2}q_{11}] + \mu_4[-p_4 + \frac{1}{2}q_{44}], \\ 0 &\geq 1 + \lambda q_{14} - \mu_4 q_{44} =: A, \\ 0 &\geq 1 + \lambda q_{11} - \mu_4 q_{14} + (\frac{\alpha_1}{m_{14}})A, \\ 0 &\geq 1 + \lambda q_{13} - \mu_3 q_{34} + (\frac{\alpha_3}{m_{34}})A. \end{aligned}$$

Similarly treating each region we obtain a linear program.

7 Incorporating virtual stations

For certain buffer priority policies, Dai *et. al* [8, 9, 10] have shown that certain regions in the state-space are transient. Indeed, some regions X^ϕ can never be entered when the initial condition $x(0)$ is the origin, which is in the single closed communicating class. This allows us to simply drop such ϕ's from the linear program, thus reducing the number of constraints. This provides a smaller upper bound and a larger lower bound on performance.

A little notation is helpful. Given a set of buffers C, let $Stat(C)$ denote the set of all stations which serve at least one buffer in C. Given a set of buffers C and a station

σ, let $Buff(C,\sigma)$ denote all those buffers in C which are served by station σ.

Definition 7.1: A virtual multiserver station. *A nonempty set of buffers C is a virtual multiserver station if*

(i) $b_1 \notin C$

(ii) $b_i \in C$ *and* $\sigma(i) = \sigma(i-1) \Rightarrow b_{i-1} \in C$

(iii) $b_i \in C$ *and* $\sigma(i) \neq \sigma(i-1) \Rightarrow b_{i-1} \notin C$ *and* $\sigma(i-1) \in Stat(C)$.

Definition 7.2: A C-Exciting Buffer Priority Policy. *A buffer priority policy will be said to excite a virtual multiserver station C if the buffers in $Buff(C,\sigma)$ have higher priority than $\{b_i \in Buff(C^c,\sigma) : b_{i+1} \in C\}$.*

The following result is due to [8, 9, 10].

Theorem 7.3. *Consider a buffer priority policy exciting a virtual multiserver station C. There is a random time $\tau < +\infty$ a.s. such that*

$$\prod_{\sigma \in Stat(C)} \left(\sum_{i \in Buff(C,\sigma)} x_i(t) \right) \quad = \quad 0 \text{ for all } t \geq \tau, \text{ a.s.}$$

The following example illustrates the above definitions and Theorems.

Example 7.4. *Consider the system shown in Figure 2. The set $C = \{b_2, b_4\}$ is a virtual multiserver station. The buffer priority policy giving priority to b_2 and b_4 excites C. Suppose $x(0) = 0$ is the initial condition. Then*

$$x_2(t)x_4(t) = 0 \qquad \text{for all } t \geq 0.$$

Thus Theorem 7.3 simply says that $x_2(t)$ and $x_4(t)$ cannot simultaneously be non-empty. To see this, suppose that they are initially empty. Consider the first time when

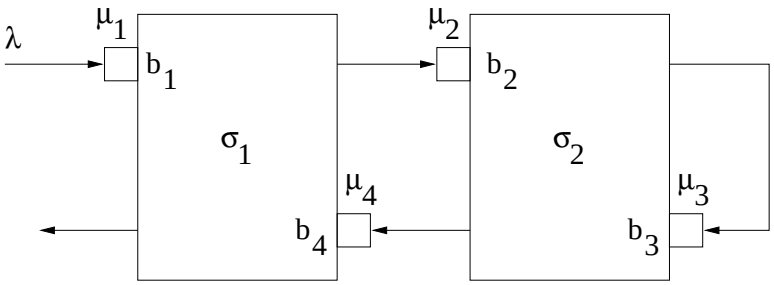

Figure 2: A network with a virtual multiserver station.

one is nonempty, say b_2. Since b_2 has priority, b_3 cannot be served. Thus b_4 cannot get any customer from b_3. Hence b_4 is empty as long as b_2 is nonempty. Now consider the first time subsequently that b_4 is nonempty. Necessarily, b_2 is empty. Since b_4 has higher priority than b_1, no customer can complete service at b_1 and move to b_2. Thus b_2 remains empty as long as b_4 is nonempty.

Continuing this argument shows that both b_2 and b_4 cannot simultaneously be nonempty. □

Suppose the buffer priority policy under consideration is C-exciting. If ϕ is such that

$$\prod_{\sigma \in Stat(C)} \sum_{i \in Buffer(C,\sigma)} \phi_i \;>\; 0, \tag{23}$$

then the region X^ϕ is transient by Theorem 7.3. Hence such a ϕ can be dropped from the linear program. This lowers the upper bound on performance, and increases the lower bound.

We note that a buffer priority policy may excite more than one virtual multiserver station. In that case we can restrict attention to those virtual multiserver stations which are *non-nested*. By this we mean a set $C_1, C_2, \ldots, C_n$ of virtual multiserver

stations for which there are no i and j satisfying the conditions:

$$i \neq j, \quad C_i \subseteq C_j \text{ and } Stat(C_i) \quad = Stat(C_j).$$

The presence of virtual multiserver stations can also be used to improve the LPs in [1] by appending some constraints. Consider first the case where there are only two stations serving a virtual multiserver station C. Let σ_1 and σ_2 denote the two stations involved. If $x_i(t) \geq 1$ for $i \in Buff(C, \sigma_1)$, then $x_j(t) = 0$ for $j \in Buff(C, \sigma_2)$. Hence $u_j(t) = 0$ also for such j. Defining $z_{ji} = E[u_j(t)x_i(t)]$ as in [1], we obtain the constraint

$$\sum_{i \in Buff(C,\sigma_1)} \sum_{j \in Buff(C,\sigma_2)} z_{ji} \quad = \quad 0. \tag{24}$$

If two or more stations serve C, then

$$\sum_{i \in C} u_i(t) \quad \leq \quad |Stat(C)| - 1.$$

Hence multiplying by any $x_j(t)$ yields the constraint

$$\sum_{i \in C} z_{ij} \quad \leq \quad |Stat(C) - 1|E[x_j(t)]. \tag{25}$$

It is easy to see that (25) is a somewhat crude way of obtaining a constraint from the action of the virtual multiserver station. It is much better to just drop a ϕ satisfying (23).

Example 7.5. *Consider the network of Figure 2, with* $\mu_1 = 1, \mu_2 = 1/2, \mu_3 = 1, \mu_4 = 1/2$, *operating under the buffer priority policy giving priority to* b_4 *and* b_2. *The virtual multiserver station* $C = \{b_2, b_4\}$ *is excited by this policy. Figure 3 provides point-wise performance bounds versus the system loading under a finite second moment*

assumption (so that one need not test whether $W(x)$ is bounded below as discussed in Appendix A.1). The dashed line bounds were obtained from the techniques of [1] (incorporating the virtual multiserver station constraints of this section). The approach of Theorem 4.1 coupled with the elimination of transient faces produces the solid line bounds. The improvement is evident. □

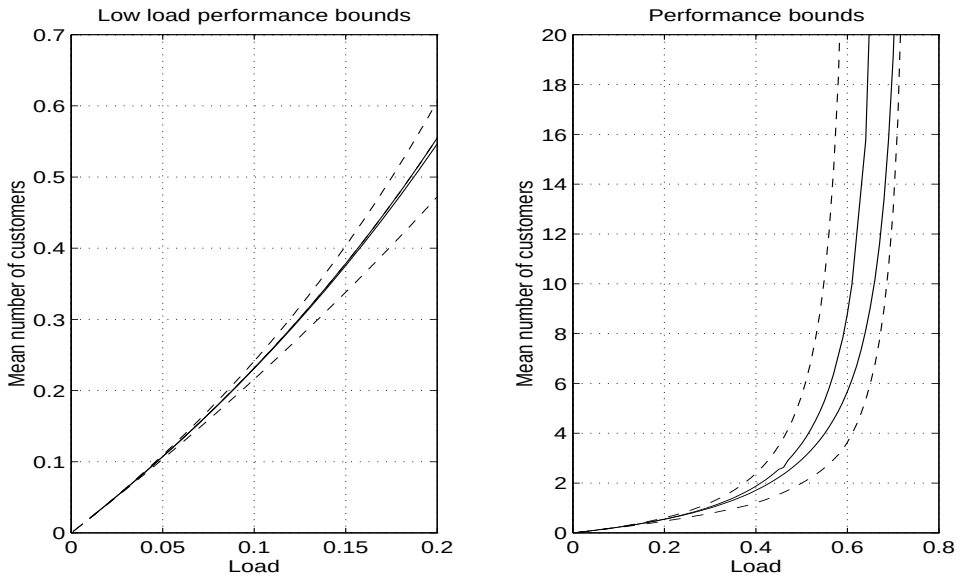

Figure 3: Performance bounds as a function of load for Example 7.5

8 Another embellishment: successive reenterancy

Here is a minor embellishment. Suppose $\sigma(i-1) = \sigma(i)$, and b_i has higher priority than b_{i-1}. Then it is easy to see that after some random time τ,

$$x_i(t) \quad = \quad 0 \text{ or } 1 \text{ for all } t \geq \tau.$$

Hence a region Y^ψ with $\psi_i = 2$ need not be considered in the linear program, again improving the bounds.

9 Functional bounds

In the preceding section, we have obtained a separate bound for each λ, that is, we have solved a separate linear program for each value of λ. Now we show, as in [5, 6], how to solve a single linear program which obtains a bound for *all* λ's.

The nominal load on the system, at an arrival rate of λ, is

$$\rho \quad := \quad \mathrm{Max}_\sigma \sum_{i \in \sigma} \frac{\lambda}{\mu_i}.$$

Thus the system can conceivably operate in the range $0 \leq \rho < 1$. However, under a particular scheduling policy the mean number in the system may be infinite at a value $\rho < 1$. To allow for this possibility we consider Laurent expansions which allow divergence at some $\rho' \leq 1$:

$$\begin{aligned} J(\rho) &= \sum_{m=-1}^{M} \frac{J_m}{(\rho' - \rho)^m} \\ c^\phi(\rho) &= \sum_{m=-1}^{M} \frac{c_m^\phi}{(\rho' - \rho)^m} \\ p^\phi(\rho) &= \sum_{m=-1}^{M} \frac{p_m^\phi}{(\rho' - \rho)^m} \\ Q^\phi(\rho) &= \sum_{m=-1}^{M} \frac{Q_m^\phi}{(\rho' - \rho)^m} \end{aligned}$$

for $0 \leq \rho < \rho'$. The quadratic expression on the region X^ϕ is just $W^\phi(x, \rho) = c^\phi(\rho) + p^{\phi T}(\rho)x + \frac{1}{2}x^T Q^\phi(\rho)x$, and the composite is

$$W(x, \rho) \quad = \quad W^\phi(x, \rho) \text{ for } x \in X^\phi.$$

For an upper bound, we would like this $J(\rho)$ and $W(x,\rho)$ to satisfy the average cost inequality below for all $0 \leq \rho < \rho'$:

$$J(\rho) - |x| \quad \geq \quad \lambda[W(x+e_1,\rho) - W(x,\rho)] + \sum_i u_i\mu_i[W(x-e_i+e_{i+1},\rho) - W(x,\rho)].$$

Let λ' be the arrival rate corresponding to the load ρ', i.e., $\rho' = \text{Max}_\sigma \sum_{i\in\sigma} \frac{\lambda'}{\mu_i}$. Then $\lambda = \frac{\lambda'}{\rho'}\rho$. Hence λ can be written as

$$\lambda \quad = \quad \lambda' - \frac{\lambda'}{\rho'}(\rho' - \rho).$$

Hence we need to satisfy

$$\begin{aligned} J(\rho) - |x| \quad \geq \quad & \lambda'[W(x+e_1,\rho) - W(x,\rho)] - \frac{\lambda'}{\rho'}(\rho'-\rho)[W(x+e_1,\rho) - W(x,\rho)] \\ & + \sum_i u_i\mu_i[W(x-e_i+e_{i+1},\rho) - W(x,\rho)]. \end{aligned}$$

Both the left-hand side and the right-hand side are expansions in powers of $(\rho'-\rho)$. We seek to satisfy the inequality by enforcing the inequalities for each of the coefficients of corresponding powers of $(\rho'-\rho)$ on both the left-hand side and right-hand side. For each power $(\rho'-\rho)^n$, we then obtain inequalities by following the method leading to Theorem 4.1, yielding to a linear program for functional bounds.

The issue of determining whether any $W(x)$ function obtained is indeed bounded below (for an upper functional bound), may be approached in two ways. The first is to require the conditions of Lemma 4.3 to hold for each power of $W^\phi(x,\rho)$, i.e., require Q_m^ϕ to be copositive and $p_m^{\phi T}x \geq 0$ for all $x \in R_+^L$ with $x^T Q_m^\phi x = 0$ for each m and ϕ. Alternately, one can write the $W(x,\rho)$ function with ρ multiplying each term and check the resulting powers for the condition, e.q., if $\rho' = 1$, require $\sum_{m=k}^M Q_m^\phi$ to be copositive for all k and similarly for the linear terms.

Regarding the choice of ρ', the obvious first choice is $\rho' = 1$. If the linear program is infeasible, then one reduces ρ' until it is feasible. Concerning M, one starts with $M = 1$. If this is infeasible, one increases it to $M = 2$ and so on until feasibility. Concerning the objective function, there is considerable latitude. If one wants to bound heavy traffic performance, then one chooses J_M. After minimizing it, there is still some latitude. One can constrain J_M at its minimum value, and then minimize J_{M-1}, and so on. Alternatively, if one wants to bound performance at a particular value $\overline{\rho}$, then the natural objective function is $\sum_{m=-1}^{M} \frac{J_m}{(\rho' - \overline{\rho})^m}$. In all cases, at feasibility one wants a gross upper bound valid for all $0 \leq \rho < \rho'$.

A similar approach also yields functional lower bounds.

The extension to upper and lower performance bounds for the class of all non-idling policies is just as in Section 5.

Example 9.1. *Consider the network of Figure 4, with* $\mu_1 = 3, \mu_2 = 3/2, \mu_3 = 3/2, \mu_4 = 3$, *operating under the LBFS (last buffer first served) buffer priority policy. In [6], the functional bounds*

$$\frac{1.2222\rho}{(1-\rho)} + 0.7778\rho \leq E \mid x(n) \mid \leq \frac{2.8333\rho}{(1-\rho)} + 0.6667\rho,$$

for all $\rho \in [0, 1)$, *were obtained. Through the approach presented here the following functional bounds were obtained,*

$$Max \left\{ \frac{1.7222}{(1-\rho)} - 1.1729 - 10.5493(1-\rho), \frac{1.4645}{(1-\rho)} - 0.9290 - 0.5355(1-\rho) \right\}$$
$$\leq E \mid x(n) \mid \leq \frac{2.8333}{(1-\rho)} - 3.6667 + 0.8333(1-\rho).$$

The lower bound improvement is clear. The coefficient of $1/(1-\rho)$ *in the upper bound remains the same but the upper bound is also an improvement. Figure 5 is a plot of the functional bounds. The dashed line bounds were obtained via the techniques of this paper, while the solid line bounds are those of [6].* □

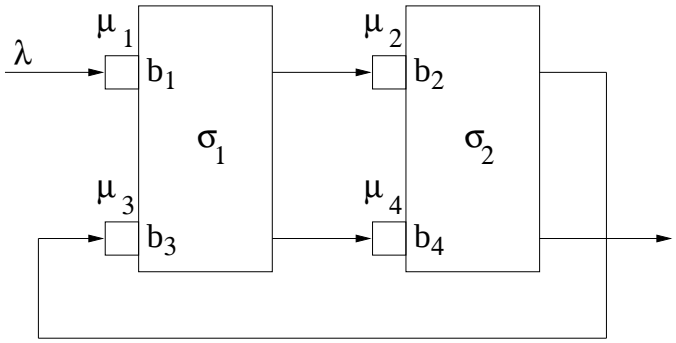

Figure 4: System of Example 9.1

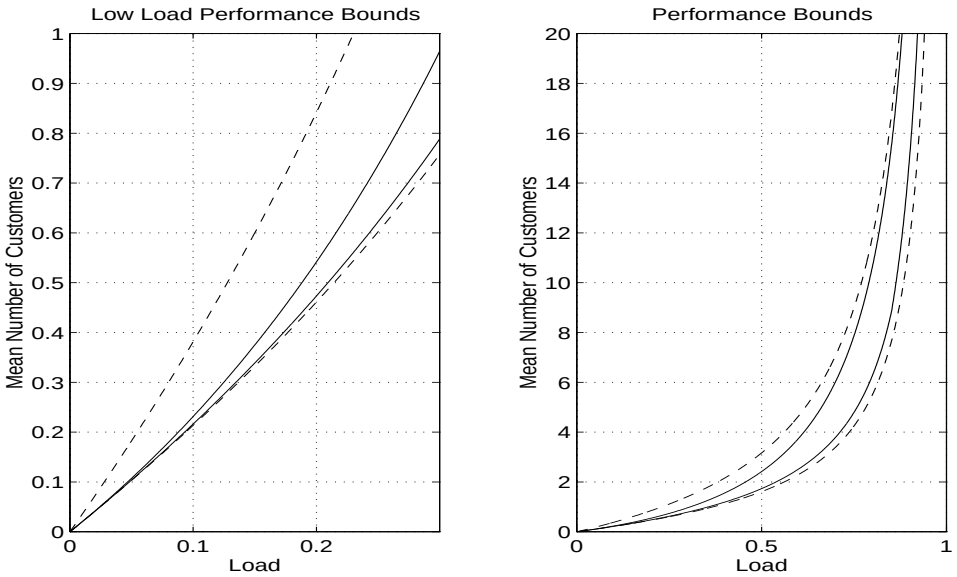

Figure 5: Functional bounds for the system of Example 9.1

10 Concluding remarks

The approach developed here exploits to the hilt the key shift invariance of transition probabilities in queueing networks. It proposes a systematic face based quadratic form for the differential costs in the average cost inequality. The net result is a provably superior approach than in [1]. However the number of ψ's to be considered can grow quickly with L, the upper bound is 3^L. The issue of size reduction of the LPs merits further attention.

References

[1] S. Kumar and P. R. Kumar, "Performance bounds for queueing networks and scheduling policies," *IEEE Transactions on Automatic Control*, vol. AC-39, pp. 1600–1611, August 1994.

[2] D. Bertsimas, I. C. Paschalidis, and J. N. Tsitsiklis, "Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance," *Annals of Applied Probability*, vol. 4, pp. 43–75, 1994.

[3] P. R. Kumar and S. P. Meyn, "Stability of queueing networks and scheduling policies," *IEEE Transactions on Automatic Control*, vol. 40, pp. 251–260, February 1995.

[4] P. R. Kumar and S. P. Meyn, "Duality and linear programs for stability and performance analysis of queueing networks and scheduling policies," *IEEE Transactions on Automatic Control*, vol. 41, pp. 4–17, January 1996.

[5] H. Jin, J. Ou, and P. R. Kumar, "The throughput of irreducible closed Markovian queueing networks: Functional bounds, asymptotic loss, efficiency, and the Harrison-Wein conjectures." To appear in *Mathematics of Operations Research*, 1997.

[6] C. Humes, Jr., J. Ou, and P. R. Kumar, "The delay of open Markovian queueing networks: Uniform functional bounds, heavy traffic pole multiplicities, and stability." To appear in *Mathematics of Operations Research*, 1997.

[7] P. R. Kumar and P. P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*. Englewood Cliffs, NJ: Prentice-Hall, 1986.

[8] J. J. Hasenbein, "Necessary conditions for global stability of multiclass queueing networks." Industrial and Systems Engineering Report Series J-96-01, May 1996. Georgia Institute of Technology.

[9] J. Dai and J. H. V. Vate, "Virtual stations and the capacity of two–station queueing networks." Preprint, March 1996. School of Industrial and Systems Engineering, Georgia Institute of Technology.

[10] J. Dai and J. H. V. Vate, "Global stability of two–station queueing networks." Preprint, 1996. School of Industrial and Systems Engineering, Georgia Institute of Technology.

[11] R. W. Cottle, G. J. Habetler, and C. E. Lemke, "On classes of copositive matrices," *Linear Algebra and Its Applications*, vol. 3, pp. 295–310, 1970.

[12] L. Andersson, G. Chang, and T. Elfving, "Criteria for copositive matrices and nonnegative Bezier patches," Tech. Rep. LiTH-MAT-R-93-27, Linkoping University and University of Science and Technology, China, August 1993.

11 Appendix A.1

We show that if one uses a single quadratic expression

$$W(x) \quad = \quad c + p^T x + \frac{1}{2} x^T Q x$$

over the entire state space, then one improves upon the LPs in [1]. This establishes the superiority of the bounds obtained from the consideration of different such W^ϕ's in different regions X^ϕ.

To show this, we must modify the assumptions to match those of [1]. Specifically, rather than requiring that $W(x)$ be bounded below, we assume a finite second moment for $x(t)$. Lemma 3.1 carries over with this change since $\frac{EW(x(N))}{N} \rightarrow 0$ under this assumption.

For specificity, we focus on a buffer priority policy θ. The result for the class of all non-idling policies is similarly established.

For a given region X^ϕ and station σ, let $b_{i(\phi,\sigma,\theta)}$ denote the buffer that is served at station σ. If all buffers at σ are empty in X^ϕ, then such a station is understood to be dropped from the summation below. The LP in Theorem 4.1 collapses to

$$\text{Min} \quad J$$

subject to

$$\begin{aligned} J \quad \geq \quad & e^T \psi + \lambda \left[p_1 + \frac{1}{2} q_{11} + \sum_k q_{1k} \psi_k \right] \\ & + \sum_\sigma \mu_{i(\phi(\psi),\sigma,\theta)} \left[p_{i(\phi(\psi),\sigma,\theta)+1} - p_{i(\phi(\psi),\sigma,\theta)} + \frac{1}{2} q_{i(\phi(\psi),\sigma,\theta)+1, i(\phi(\psi),\sigma,\theta)+1} \right. \\ & \left. + \frac{1}{2} q_{i(\phi(\psi),\sigma,\theta), i(\phi(\psi),\sigma,\theta)} - q_{i(\phi(\psi),\sigma,\theta), i(\phi(\psi),\sigma,\theta)+1} \right. \end{aligned}$$

$$+\sum_k \left(q_{i(\phi(\psi),\sigma,\theta)+1,k} - q_{i(\phi(\psi),\sigma,\theta),k}\right)\psi_k\Bigg],$$

$$1 + \lambda q_{1j} + \sum_\sigma \mu_{i(\psi,\sigma,\theta)}(q_{i(\phi(\psi),\sigma,\theta)+1,j} - q_{i(\phi(\psi),\sigma,\theta),j}) \leq 0, \forall \text{ j such that } \psi_j = 2,$$

for all ψ.

The second constraint above is the same as the constraint of the duals of the original LPs, [see [4]]. Take any Q satisfying them. The choice $p_i = \frac{1}{2}q_{ii} + \sum_{k=i+1}^{L} q_{k,k} - \sum_{k=1}^{L} q_{k,k+1}$ is then feasible. This coincides with the objective function of the duals of the original LPs, see [4]. This establishes that the value is at least as good as that of the original LPs. The proof for the lower bound is similar.

The proof of superiority extends when virtual multiserver station constraints (24) and (25) are added to the constraints of [1], while the ϕ's satisfying (23) are removed as in Section 6. Actually in this case the incorporation of virtual multiserver stations through the elimination of certain ϕ's is superior to the somewhat crude consequences of virtual multiserver stations captured by (24,25).

12 Appendix A.2: The definitions of d^ψ, r^ψ and S^ψ

Let $e = (1, \ldots, 1)^T$. Note that $e^T x = |x|$ for $x \in Z_+^L$. Then

$$\begin{aligned} d^\psi \quad := \quad & e^T\psi + \lambda\Big[c^{\phi(\psi+e_1)} - c^{\phi(\psi)} + p_1^{\phi(\psi+e_1)} + \frac{1}{2}q_{11}^{\phi(\psi+e_1)} \\ & +(p^{\phi(\psi+e_1)} - p^{\phi(\psi)} + Q^{\phi(\psi+e_1)}e_1)^T\psi + \frac{1}{2}\psi^T(Q^{\phi(\psi+e_1)} - Q^{\phi(\psi)})\psi\Big] \\ & + \sum_{i:u_i(\phi(\psi))=1} \mu_i \Big[c^{\phi(\psi-e_i+e_{i+1})} - c^{\phi(\psi)} - p_i^{\phi(\psi-e_i+e_{i+1})} \\ & p_{i+1}^{\phi(\psi-e_i+e_{i+1})} + \frac{1}{2}q_{ii}^{\phi(\psi-e_i+e_{i+1})} + \frac{1}{2}q_{i+1,i+1}^{\phi(\psi-e_i+e_{i+1})} - \end{aligned}$$

$$q_{i,i+1}^{\phi(\psi-e_i+e_{i+1})} + \left(p^{\phi(\psi-e_i+e_{i+1})} + Q^{\phi(\psi-e_i+e_{i+1})}(e_{i+1}-e_i) - p^{\phi(\psi)}\right)^T \psi$$
$$\frac{1}{2}\psi^T(Q^{\phi(\psi-e_i+e_{i+1})} - Q^{\phi(\psi)})\psi\Big],$$

$$r^\psi := e + \lambda\left[p^{\phi(\psi+e_1)} + Q^{\phi(\psi+e_1)}(e_1+\psi) - p^{\phi(\psi)} - Q^{\phi(\psi)}\psi\right]$$
$$+ \sum_{i:u_i(\phi(\psi))=1} \mu_i\left[p^{\phi(\psi-e_i+e_{i+1})} + Q^{\phi(\psi-e_i+e_{i+1})}(\psi+e_{i+1}-e_i) - p^{\phi(\psi)} - Q^{\phi(\psi)}\psi\right],$$

$$S^\psi := \lambda\left[Q^{\phi(\psi+e_1)} - Q^{\phi(\psi)}\right] + \sum_{i:u_i(\phi(\psi))=1} \mu_i\left[Q^{\phi(\psi-e_i+e_{i+1})} - Q^{\phi(\psi)}\right].$$