Optimal Buffer Allocation in Tandem Queues with Communication Blocking

Dong-Won Seo, Sung-Seok Ko, and Uk Jung

ABSTRACT—In this letter, we consider an m-node tandem queue (queues in series) with a Poisson arrival process and either deterministic or non-overlapping service times. With the assumption that each node has a finite buffer except for the first node, we show the non-increasing convex property of stationary waiting time with respect to the finite buffer capacities. We apply it to an optimization problem which determines the smallest buffer capacities subject to probabilistic constraints on stationary waiting times.

Keywords—Buffer allocation, (max,+)-algebra, (max,+)-linear system, tandem queue, timed event graphs, waiting times.

I. Introduction

As a common model of telecommunication networks, finite or infinite-buffer tandem queues have been widely studied. However, because of finiteness, most studies over the past decades focused on very restrictive systems. Recently, various types of networks belonging to a more generous type of system, the so called (max, +)-linear system, have been studied. They can be properly modeled by timed event graphs. A (max, +)-linear system is a choice-free network of single-server queues with FIFO service discipline.

This study aims to show the non-increasing convex property of waiting time characteristics derived in [1] and to apply it to an optimization problem determining the smallest buffer capacities satisfying probabilistic constraints on stationary waiting times. To the best of our knowledge, there has been no result providing a mathematical proof for stochastic network models. Moreover, it is useful to manipulate various cost-time related functions. Refer to [2] for basic (max, +)-algebra and some preliminaries on waiting times in a (max, +)-linear system.

II. Preliminaries and Main Result

The basic reference algebra throughout this study is the so called (max, +)-algebra on the real line $\mathbb{R}$, namely, the semi-field with the two operations ($\oplus$, $\otimes$), where $\oplus$ means maximization and $\otimes$ means addition for scalars and the (max, +)-algebra product for matrices (see [2]). The dynamics (stochastic behaviors) of (max, +)-linear systems can be captured by the following α-dimensional vectorial recurrence equation:

$$X_{n+1} = A_n \otimes X_n \oplus B_{n+1} \otimes T_{n+1} \tag{1}$$

with an initial condition X_0, where $\{T_n\}$ is a non-decreasing sequence of real-valued random numbers (the epochs of the Poisson arrival process with rate λ); $\{A_n\}$ and $\{B_n\}$ are stationary and ergodic sequences of real-valued random matrices of size $\alpha\times\alpha$ and $\alpha\times1$, respectively; and $\{X_n\}$ is a sequence of α-dimensional state vectors. The components of the state vector represent absolute times which grow to ∞ when n increases unboundedly; hence, we are more interested in the differences $W_n^i = X_n^i - T_n$ (like the waiting time of the n-th customer before joining server i). Let $\tau_n = T_{n+1} - T_n$ with T_0=0, and let $C(x)$ be the $\alpha\times\alpha$ matrix with all diagonal entries equal to $-x$ and all non-diagonal entries equal to $-\infty$. By subtracting T_{n+1} from both sides of (1), the new state vector W_{n+1} can be written as

$$W_{n+1} = A_n \otimes C(\tau_n) \otimes W_n \oplus B_{n+1}$$

Manuscript received July 28, 2008; revised Sept. 27, 2008; accepted Oct. 21, 2008.

This research was supported by the Kyung Hee University Research Fund in 2008 (KHU-20080632).

Dong-Won Seo (phone: + 82 31 201 2311, e-mail: dwseo@khu.ac.kr,) is with the College of Management and International Relations, Kyung Hee University, Yongin, Rep. of Korea.

Sung-Seok Ko (e-mail: ssko@konkuk.ac.kr) is with the Department of Industrial Engineering, Konkuk University, Seoul, Rep. of Korea.

Uk Jung (e-mail: ukjung@dongguk.edu) is with the Department of Management, Dongguk University, Seoul, Rep. of Korea.

for $n \geq 0$ and with some initial condition W_0. Under certain conditions, it is shown in [2] and in theorem 1 of [3] that for all $\lambda < a^{-1}$, where a is the maximal Lyapunov exponent, the stationary waiting time W is determined by the matrix-series

$$W = D_0 \oplus \bigoplus_{k \geq 1} C(-T_{-k}) \otimes D_k \tag{2}$$

with D_0=B_0 and W_0=B_0, and for all $k \geq 1$,

$$D_k = \left(\bigotimes_{n=1}^{k} A_{-n} \right) \otimes B_{-k} . \tag{3}$$

Using this topology in [4] and [3], it was possible to obtain waiting time characteristics in a class of stochastic networks as a Taylor series expansion with respect to an arrival rate λ. Note that the random vector D_n plays an important role in computing waiting time characteristics, and can be calculated independently of the arrival rate. In addition, the components of D_n can be interpreted as a critical path in a task graph and written in terms of service times.

From the definition of D_n in (3) with some algebra, we can derive the explicit expressions of D_n for a (max, +)-linear system. In [1], they derived the explicit expression of D_n^i, the i-th component of D_n, for stationary waiting times in either constant (the following proposition 1) or non-overlapping finite-buffer tandem queues with communication blocking. Under a communication blocking policy, a customer at node j cannot begin to use a service unless there is a vacant space in the buffer at node $j+1$. For a node i, let σ^i be the constant service time, and let K_i be the finite buffer size including room for a customer to be served, and denote $D_n^i\left(\mathbf{K}^{i+1}\right)$ as a D_n^i with finite buffers $\mathbf{K}^{i+1}$=$\left(K_{i+1}, \cdots, K_m\right)$.

Proposition 1. In a constant m-node tandem queue with communication blocking with the assumptions that $K_j \geq 2$, $j = 2, \cdots, m$, and $K_1 = K_{m+1} = \infty$, $D_n^i\left(\mathbf{K}^{i+1}\right)$ is given as

$$D_n^i\left(\mathbf{K}^{i+1}\right) = \sum\nolimits_{j=1}^{i-1} \sigma^j + \max\left\{n\sigma^1, \cdots, n\sigma^i\right\} \text{ for } 0 \leq n < K_{i+1}, \tag{4}$$

$$D_n^i\left(\mathbf{K}^{i+1}\right) = \sum\nolimits_{j=1}^{i-1} \sigma^j + \max\left\{n\sigma^1, \cdots, n\sigma^i, \ell_i(i+1), \cdots, \ell_i(\kappa_i)\right\}$$

$$\text{for } \sum\nolimits_{j=i+1}^{\kappa_i} K_j \leq n < \sum\nolimits_{j=i+1}^{\kappa_i+1} K_j , \tag{5}$$

where $\ell_i(p) = \sigma^i + 2\sum\nolimits_{j=i+1}^{p-1} \sigma^j + \left[n - \left(\sum\nolimits_{j=i+1}^{p} K_j \right) + 1 \right] \sigma^p$,

$$\kappa_i = \min\left\{ q \in (i+1, \cdots, m) : \sum\nolimits_{j=i+1}^{q} K_j \leq n < \sum\nolimits_{j=i+1}^{q+1} K_j \right\}$$

with the convention that summation over an empty set is 0.

From this result, theorem 1 follows, which shows the non-increasing convex property of $D_n^i\left(\mathbf{K}^{i+1}\right)$ in finite buffer $\mathbf{K}^{i+1}$.

Theorem 1. In either a constant or non-overlapping m-node tandem queue with communication blocking, $D_n^i\left(\mathbf{K}^{i+1}\right)$ is non-increasing convex in each finite buffer of $\left(K_{i+1}, \cdots, K_m\right)$ and is more sensitive to the closer buffer capacity among the downstream nodes. That is, for all $n \geq 0$,

$$\begin{aligned} D_n^i\left(\mathbf{K}^{i+1} + e_{i+1}\right) \leq D_n^i\left(\mathbf{K}^{i+1} + e_{i+2}\right) \leq \cdots \\ \cdots \leq D_n^i\left(\mathbf{K}^{i+1} + e_m\right) \leq D_n^i\left(\mathbf{K}^{i+1}\right), \end{aligned} \tag{6}$$

where $e_j = \left(\overbrace{0, \cdots, 0}^{j-(i+1)}, 1, \underbrace{0, \cdots, 0}_{m-j} \right)$.

Proof. Consider a deterministic tandem queue first. Assume that we focus on a node i, $i \in \{1, \cdots, m\}$, and that a buffer at node k, $k \in \{i+1, \cdots, m\}$, is increased by one and the other buffers remain the same, denoted by $\mathbf{K}^{i+1} + e_k$. Equations (4) and (5) can be written as follows. When k=i+1,

$$D_n^i\left(\mathbf{K}^{i+1} + e_k\right) = \sum\nolimits_{j=1}^{i-1} \sigma^j + \max\left\{n\sigma^1, \cdots, n\sigma^i\right\} \text{ for } 0 \leq n \leq K_i ,$$

$$D_n^i\left(\mathbf{K}^{i+1} + e_k\right) = \sum\nolimits_{j=1}^{i-1} \sigma^j + \max\left\{n\sigma^1, \cdots, n\sigma^i, \hat{\ell}_i(i+1), \cdots, \hat{\ell}_i(\nu_i)\right\}$$

$$\text{for } \sum\nolimits_{j=i+1}^{\nu_i} K_j < n \leq \sum\nolimits_{j=i+1}^{\nu_i+1} K_j , \tag{7}$$

where $\hat{\ell}_i(p) = \sigma^i + 2\sum\nolimits_{j=i+1}^{p-1} \sigma^j + \left[n - \left(\sum\nolimits_{j=i+1}^{p} K_j \right) \right] \sigma^p$,

$$\nu_i = \min\left\{ q \in (i+1, \cdots, m) : \sum\nolimits_{j=i+1}^{q} K_j < n \leq \sum\nolimits_{j=i+1}^{q+1} K_j \right\} .$$

Similarly, when $k \geq i+2$,

$$D_n^i\left(\mathbf{K}^{i+1} + e_k\right) = \sum\nolimits_{j=1}^{i-1} \sigma^j + \max\left\{n\sigma^1, \cdots, n\sigma^i\right\} \text{ for } 0 \leq n < K_i ,$$

$$D_n^i\left(\mathbf{K}^{i+1} + e_k\right) = \sum\nolimits_{j=1}^{i-1} \sigma^j + \max\left\{n\sigma^1, \cdots, n\sigma^i, \tilde{\ell}_i(i+1), \cdots, \tilde{\ell}_i(\omega_i)\right\}$$

$$\text{for } \begin{cases} \sum\nolimits_{j=i+1}^{\omega_i} K_j \leq n < \sum\nolimits_{j=i+1}^{\omega_i+1} K_j & \text{if } \omega_i < k-1 \\ \sum\nolimits_{j=i+1}^{k-1} K_j \leq n \leq \sum\nolimits_{j=i+1}^{k} K_j & \text{if } \omega_i = k-1 , \\ \sum\nolimits_{j=i+1}^{\omega_i} K_j < n \leq \sum\nolimits_{j=i+1}^{\omega_i+1} K_j & \text{if } \omega_i \geq k-1 \end{cases} \tag{8}$$

where $\omega_i = \kappa_i$ if $\omega_i < k-1$ and $\omega_i = \nu_i$ if $\omega_i \geq k$, and $\tilde{\ell}_i(\cdot) = \ell_i(\cdot)$ if $\omega_i \leq k$ and $\tilde{\ell}_i(\cdot) = \hat{\ell}_i(\cdot)$ if $\omega_i > k$, in which κ_i, ν_i, $\ell_i(\cdot)$, and $\hat{\ell}_i(\cdot)$ are the same as those previously defined.

Now, we can see that $\hat{\ell}_i(\cdot) \leq \ell_i(\cdot)$ for any n with a given $\mathbf{K}^{i+1}$, and $\hat{\ell}_i(\cdot)$ in (7) is replaced by a $\ell_i(\cdot)$ from the left to the right as k is increased by one. Thus, (8) can have the non-increasing property in k. Then, from the definition of D_n^i (see (3)) and the linearity and convexity of the "max" function, we can infer the fact given in (6). Moreover, the same arguments are also valid for a tandem queue with non-overlapping service times (see proposition 2 in [1]), which completes the proof. □

This non-increasing property of $D_n^i\left(\mathbf{K}^{i+1}\right)$ and the fact that the composition of convex functions is also convex (see (2)) imply that W^i, which is the elapsed time from the arrival until the beginning of service at node i, is also a non-increasing convex in $\left(K_{i+1},\cdots,K_m\right)$.

III. Application and Examples

For a node *i*, let $\tau_i \ge 0$ be a pre-specified bound on waiting time W^i and let $0 < \beta_i < 1$ be a pre-specified probability value, such as QoS. Because the system sojourn time W^m, that is, the waiting time at the last node m, is independent of finite buffer capacities in either constant or non-overlapping service times (see [5]), we consider only sub-areas of the system. For a simple instance, the optimal buffer capacities can be computed as the solution of the following optimization problem. For a given arrival rate $\lambda \in [0, a^{-1})$, where $a = \max\left\{\sigma^1,\cdots,\sigma^m\right\}$,

$$\begin{aligned} &\min \quad \sum_{i=2}^{m} K_i \\ &s.t. \quad \Pr(W^i > \tau_i) \le \beta_i \quad \text{for } i = 1,\cdots,m-1. \\ &\qquad\; K_i \in \mathbf{N} \end{aligned}$$

Our main results, the non-increasing convex properties of $D_n^i\left(\mathbf{K}^{i+1}\right)$ and W^i, guarantee the existence of optimal solutions of this problem. By using this fact together with an explicit expression of $D_n^i\left(\mathbf{K}^{i+1}\right)$ given in [1] and a closed-form expression for the tail probability of stationary waiting time given in theorem 2.3 of [4], we can numerically determine optimal buffer capacities. Moreover, this optimization problem can be separately solved in reverse order of *i* one by one from $i = m-1$ to i=2 because $D_n^i\left(\mathbf{K}^{i+1}\right)$ is a function of $\left(K_{i+1},\cdots,K_m\right)$. That is to say, one can first choose the optimal value of K_m^* and then choose the optimal value of $K_{m-1} + K_m^*$ by using this value (determined just before) of K_m^*, and so on.

Our results are valid for both constant and non-overlapping service times, but to avoid computational complexity we consider a 5-node tandem queue with deterministic service times. Let $\sigma^i = 0.1 \times i$ be a constant service time at node *i*. Table 1 shows tail probabilities of waiting times at node 3 when the traffic intensity $\rho = 0.9$ with varying K_4 and K_5, which infers the non-increasing convex property of W^3 mentioned in theorem 1 (see the shaded cells). Table 2 shows the optimal buffer sizes satisfying probabilistic constraints on waiting times when $\tau_1 = 0.5, \tau_2 = 1.5, \tau_3 = 3.0,$ and $\tau_4 = 4.5$.

Table 1. Pr(W^3>0.5) for various K_4 and K_5.

K_5 K_4	2	3	4	5	6
2	0.754152	0.692997	0.653213	0.624306	0.602422
3	0.662715	0.600935	0.557714	0.525635	0.501105
4	0.598502	0.54857	0.510253	0.481023	0.458369
5	0.54857	0.508774	0.476992	0.452034	0.43242
6	0.508774	0.47687	0.451126	0.430459	0.41401

Table 2. Optimal buffer allocation.

	K_2^*	K_3^*	K_4^*	K_5^*
$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.15$	2	4	4	3
$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.13$	3	3	4	4
$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.11$	3	3	4	5
$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.09$	3	3	4	6
$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.07$	2	4	4	7

IV. Concluding Remark

In finite-buffer *m*-node tandem queues with communication blocking, we showed the mathematical proofs for the non-increasing convex properties of $D_n^i\left(\mathbf{K}^{i+1}\right)$ and W^i with respect to finite buffers. These properties are immediately applicable to manipulate various time related functions. Moreover, these analytic methods can be extended to more complex (max, +)-linear systems such as fork-and-join networks with various blocking policies, tandem queues without the exception of finite buffer capacity at the first node, Kanban systems, and so on.

References

[1] D.-W. Seo, H. Lee, and S.-S. Ko, “Stationary Waiting Times in m-Node Tandem Queues with Communication Blocking,” *Int'l J. Manage. Sci.*, vol. 14, no. 1, 2008, pp. 23-34.

[2] F. Baccelli et al., *Synchronization and Linearity: An Algebra for Discrete Event Systems*, John Wiley & Sons, 1992.

[3] F. Baccelli and V. Schmidt, “Taylor Series Expansions for Poisson Driven (Max,+) Linear Systems,” *Annals of Applied Probability*, vol. 6, no. 1, 1996, pp. 138-185.

[4] H. Ayhan and D.-W. Seo, “Tail Probability of Transient and Stationary Waiting Times in (Max,+)-Linear Systems,” *IEEE Trans. Automatic Control*, vol. 47, no. 1, 2002, pp. 151-157.

[5] Y.-W. Wan and R.W. Wolff, “Bounds for Different Arrangements of Tandem Queues with Nonoverlapping Service Times,” *Management Science*, vol. 39, no. 9, 1993, pp. 1173-1178.