Markov Chain Decomposition Based On Total Expectation Theorem

Katsunobu Sasanuma
College of Business, Stony Brook University, Stony Brook, NY 11794, katsunobu.sasanuma@stonybrook.edu

Robert Hampshire
Gerald R. Ford School of Public Policy, University of Michigan, Ann Arbor, MI 48109, hamp@umich.edu

Alan Scheller-Wolf
Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213, awolf@andrew.cmu.edu

Markov chain (MC) decomposition methods comprise two main steps: (1) a procedure to decompose the full MC, and (2) a procedure to aggregate the solutions of decomposed subsets. While a variety of unique techniques have been applied to accomplish the first step, a normalization procedure is almost always used in the second step. The normalization does, indeed, unify disjoint subsets and retrieve the properties of the full MC. However, use of normalization makes a divide-and-conquer approach costly, and obscures the relationship between decomposed subsets and the full MC. Here we develop a novel MC decomposition method based on the total expectation theorem. Thanks to the linear property of the total expectation theorem, our method doesn’t explicitly require a normalization condition and hence simplifies the second step: The procedure to obtain the properties of the full MC is a simple summation of the properties of decomposed subsets. Moreover, unlike other methods, our method allows us to use overlapping or nested subsets, whose dependencies are maintained using a general termination scheme (modification of transition rates at boundary states after decomposition). To illustrate our method, we apply it to several MCs, decompose them into overlapping or nested subsets, and derive previously unknown properties.

Key words: Markov chain, queueing system, total expectation theorem, law of total expectation, decomposition, partial flow balance, partial flow conservation, termination, truncation.

1. Introduction

A divide-and-conquer approach is commonly used to find a quantity of interest in a complex system. This approach is generally desirable because a “complex” system is often composed of multiple simpler subsystems, each of which may be easy to analyze in isolation. In fact, a variety of divide-and-conquer methods are used in many fields including Markov chains (MCs). [Stewart (1994)] reviews major MC decomposition methods that have been developed and utilized. These methods comprise two main steps: (1) decomposition/disaggregation procedure, and (2) composition/aggregation procedure. The first step is to decompose a full MC into smaller, disjoint subsets...
(subchains). Many decomposition techniques have been developed to analyze specific MC structures. Some create a set of almost independent (nearly completely decomposable) subchains in order to find an approximate solution for the full MC; others partition a transition matrix representing a MC and solve a linear system of equations using matrix techniques. On the other hand, the second step is fundamentally the same: A normalization condition is imposed to combine the properties of the subchains and obtain the properties of the full MC. This normalization condition is absolutely necessary and is taken for granted in dealing with MCs. However, due to the normalization condition the whole is not a simple summation of parts - we need to ensure that weights of disjoint subsets sum up to one. This normalization condition is a simple requirement, but it is sufficient to discourage many people from using a MC decomposition method. For example, many queueing systems (e.g., M/M/k) can be represented by a combination of two simpler systems, but no textbooks solve these problems using MC decomposition methods; the benefits of using a MC decomposition method may not compensate for the cost of aggregating solutions of subchains. In addition, the normalization condition obscures the relationship between subchains and the full MC.

The method we propose takes a new divide-and-conquer approach for MCs. Our method is based on the total expectation theorem (the law of total expectation), which allows us to obtain the property of the whole by summing up the properties of the parts. Thanks to the linear property of the total expectation theorem, we can not only calculate a quantity of interest with ease, but also reveal how each subchain contributes to the quantity of interest of the full system. In addition, we can choose subchains with more flexibility: We do not require decomposed subchains to be disjoint. If subchains overlap with each other, we simply need to subtract the contribution from the intersecting sets. For nested subsets (i.e., each subchain is a superset/subset of other subchains), we adjust the solution by adding or subtracting the contribution from the differential sets. Finally, our method is as general as the total expectation theorem: A quantity of interest can be calculated following our method as long as it is expressed as an expectation. Our method makes it easy to aggregate the properties of subchains, and greatly reduces the burden of the second step of MC decomposition methods.

Over the past several decades, various MC decomposition methods have been developed. For the simplest example, if a detailed (or partial) flow balance condition holds for a MC, this MC is called reversible (or quasi-reversible), and any truncated subchains maintain the steady-state probability distribution of the full MC (up to a normalization constant). Hence, properties of these truncated subchains can be analyzed independently (see Whittle 1986 and Kelly 1979). Once properties of subchains are solved independently, the results are combined using a normalization condition to obtain the properties of the full MC. If a MC is “lumpable” (Kemeny and Snell 1960), then the MC can be partitioned into multiple subchains that maintain the steady-state probability distribution
of the full MC. The method of lumping has been extended by many researchers (see, for example, [Stewart 1994]). In a special case where a subchain has a single input state from other subchains, a MC is called a Single-Input Superstate Decomposable Markov Chain (SISDMC), which makes the lumping procedure simpler (see [Feinberg and Chiu 1987]). This lumping procedure can be repeatedly applied to simplify the analysis of MCs ([Katehakis and Smit 2012]).

Another MC decomposition approach is the method specifically developed for a nearly completely decomposable (NCD) MC ([Simon and Ando 1961]). Under the NCD condition, a large MC is clustered into a small number of subchains, each of which is relatively independent from other subchains. By introducing a coupling matrix representing the transitions among these subchains, an approximate solution of the full MC is efficiently calculated. Lastly, if we need to analyze a MC with a general structure using a divide-and-conquer approach, recursive algorithms are necessary. One of the most popular methods is the iterative aggregation/disaggregation (IAD) algorithm by [Takahashi 1975], which is suitable for solving large MCs numerically. This method repeatedly decomposes a full MC into partitions (disjoint subchains) and aggregates the solutions of subchains until the equilibrium is reached. The IAD method has been further extended in subsequent literature (see [Stewart 1994] for various algorithms).

Many MC decomposition methods have developed unique techniques to cope with various specific MC structures. However, all of these methods require a normalization condition when aggregating the properties of subchains; hence, these subchains must be disjoint, otherwise a normalization condition cannot be imposed. Consequently, traditional methods cannot utilize overlapping or nested subchains. Our method works for any collectively exhaustive subchains of the full MC – we develop a decomposition procedure that is not limited to a disjoint decomposition.

In general, to make a divide-and-conquer approach work in MC settings one must maintain the correct dependencies among subchains, whatever their structure (overlapping, nested, or disjoint). For this purpose, we generalize the partial flow balance condition ([Whittle 1986] and [Kelly 1979]) under which a truncated subchain can maintain the same steady-state distribution as in the original full MC, up to a normalization constant. In our generalization, rather than requiring partial flow balance to hold, we require that any partial flow imbalance be conserved at all “boundary” states identified in the decomposition process. The conservation of these unbalanced flows is made possible by adding new transitions among the boundary states of subchains, which consequently preserves the interdependence among subchains after decomposition. We call the addition of these transitions “termination.” For any sets of appropriately terminated decomposed subchains, regardless of whether they are overlapping or not, the total expectation theorem can then be straightforwardly applied. In theory, we can apply this method to any MC. In practice, however, the method requires finding a correct termination, which may be cumbersome. While in
many cases this may require only a simple procedure, finding correct terminations will, in some cases, require a recursive solution of linear equations.

To summarize, the benefits of our method are: (1) It is an exact method that explicitly takes the interdependence of subchains into account, allowing subchains to overlap or to be nested; (2) It can be applied to obtain expectations of any function, including steady-state probabilities as a special case; and (3) It satisfies an additivity property for unions of subchains, which requires no complex post-processing procedure. This final property is especially beneficial for both analytical and numerical evaluation, because a performance indicator of the full system is obtained simply from the sum of the subchain results, each of which is often easy to obtain. As this additivity property implies, our method is essentially a change of measure: We first convert the standard probability measure to a new measure that holds a desirable additivity property in queueing systems (or MCs), then evaluate our new measure, and finally convert it back to the standard probability measure.

The rest of the paper is organized as follows: In Section 2, we explain the idea of conservation of steady-state distribution – maintaining the proportionality of the steady-state distribution of a subchain in isolation to its stationary distribution in the original MC – and prove the total expectation theorem in MC settings under the conservation of distribution condition. A variant of this theorem indicates that conservation of steady-state distribution guarantees there exists a measure which ensures the existence of a simple decomposition process for any sets of subchains that compose a MC. In Section 3, we discuss how we can ensure conservation of steady-state distribution. Specifically, we first show a necessary and sufficient condition to conserve a distribution after decomposition: A boundary condition. This boundary condition is then equivalently expressed by our partial flow conservation condition, which can be satisfied by correctly imposing termination. Finally, we show several schemes guaranteed to generate appropriate termination, ensuring partial flow conservation (and consequently, conservation of distribution). In Section 4, we consider examples that reveal the advantages of our method: By applying our method to overlapping or nested subchains of the full MC, we can derive relationships among performance measures. We also apply our method to a queueing system, in which the service rate depends on the level of congestion. This example has heretofore been difficult to solve, but using our method we can derive analytical solutions by summing up the analytical solutions of terminated subchains. In the last section, we summarize the key ideas of our method and conclude the paper with a brief discussion of some of the other potential applications of our method.
2. Total Expectation Theorem for a MC

2.1. Preliminaries

We consider an ergodic (i.e., positive recurrent) continuous time MC that we decompose into multiple subchains, each of which is indexed with $j$, $j \in J^+$. We denote the set of states that compose a subchain $j$ as $A_j$ and the whole set of states in the full MC as $S$. These decomposed subchains $\{A_j : j \in J^+\}$ should be collectively exhaustive (i.e., a collection of decomposed subchains should form the full MC: $S = \bigcup_{j \in J^+} A_j$), but not necessarily mutually exclusive (disjoint). In order to correct excess contribution from overlapping states in $\{A_j : j \in J^+\}$ when calculating a total expectation, we augment the set of subchains $\{A_j : j \in J^+\}$ using another set of subchains $\{A_j : j \in J^-\}$ that satisfies the following condition, where $I(\cdot)$ represents the indicator function:

$$\sum_{j \in J^+} I(k \in A_j) - \sum_{j \in J^-} I(k \in A_j) = 1, \forall k \in S. \quad (1)$$

By considering all subchains in $J = J^+ \cup J^-$ (i.e., by adding up states in $\{A_j : j \in J^+\}$ and subtracting states in $\{A_j : j \in J^-\}$), we can make each state in $S$ contribute to the total expectation exactly once. Note that if $\{A_j : j \in J^+\}$ forms a partition of the full set $S$, the index set $J^-$ becomes a null set; however, a set $J^+$ cannot be a null set. Note also that $A_j, j \in J$, can be a subchain that is composed of a single state, and that $A_j$ may refer to the same subchain as $A_i$ with a different index $i(\neq j) \in J$.

We denote stationary distributions of the full MC and a subchain $j$ as $\pi_k^S, \forall k \in S$, and $\pi_k^{A_j}, \forall k \in A_j$, respectively. We also denote expectations evaluated on the full MC and on a subchain $j$ as $E_S[\cdot]$ and $E_{A_j}[\cdot]$, respectively. In this paper we use the following simpler notations: $\pi_k \doteq \pi_k^S$, $\pi_k^j \doteq \pi_k^{A_j}$, $E[\cdot] \doteq E_S[\cdot]$, and $E_j[\cdot] \doteq E_{A_j}[\cdot]$.

Throughout the paper, we consider the case in which a decomposed subchain maintains the same stationary distribution as the full MC (up to a normalization constant): $\pi_k \propto \pi_k^j, \forall k \in A_j, \forall j \in J$. We call this condition conservation of distribution. Obviously, a stationary distribution of a subchain is strongly affected by how we decompose the full MC and how we treat transition rates that are lost through the decomposition procedure. Our procedure that ensures conservation of the distribution of a decomposed subchain is discussed later in this paper.

The conservation of distribution condition, or simply conservation of distribution, can be expressed in several different ways. To show certain conditions equivalent to conservation of distribution, we define the following quantities:

**Definition 1.** We define the following quantities that are functions of $\pi_k, \forall k \in S$:

1. The ratio of steady-state probabilities:

$$\beta_{kk'} \doteq \frac{\pi_k}{\pi_{k'}}.$$
The conditional steady-state probability (of the original MC) of being in a state \( k \) given it is in a subchain \( j \):
\[
p_j^k = \pi_k \sum_{k' \in A_j} \pi_k',
\]

The probability of being in a subchain \( j \):
\[
P_j = \sum_{k \in A_j} \pi_k.
\]

The conditional expectation of a function \( f(\cdot) \) given that the MC is in a subchain \( j \):
\[
E[f(X)|A_j] = \sum_{k \in A_j} f(k)p_j^k.
\]

Notice that these quantities are functions of \( \pi_k \), not \( \pi_j^k \). The conditions equivalent to conservation of distribution connect quantities defined by \( \pi_k \) with functions of \( \pi_j^k \), as we see in the following proposition:

PROPOSITION 1. (Conservation of distribution)

A stationary distribution is conserved for a subchain if and only if any of the following equivalent conditions are satisfied:

(a)
\[
\pi_k \propto \pi_j^k, \forall k \in A_j, \forall j \in J.
\]

(a')
\[
\beta_{kk'} = \frac{\pi_j^k}{\pi_j^{k'}}, \forall k \in A_j, \forall k' \in A_j, \forall j \in J.
\]

(b)
\[
p_j^k = \pi_j^k, \forall k \in A_j, \forall j \in J.
\]

(b')
\[
P_j = \frac{\pi_j}{\pi_j^k}, \forall k \in A_j, \forall j \in J.
\]

(c)
\[
E[f(X)|A_j] = E_j[f(X)], \forall j \in J \text{ and for any function } f(X).
\]

Proof of Proposition 1. Condition (a), conservation of distribution, can be expressed as the equivalent condition (a') using \( \beta_{kk'} \), the ratio of steady-state probabilities. Also, condition (b) can be expressed by the equivalent condition (b') using the definitions \( p_j^k = \pi_k / \sum_{k' \in A_j} \pi_k' \) and \( P_j = \sum_{k \in A_j} \pi_k \). We prove the equivalence of the remaining conditions below:
(a’) ⇒ (b): \( \forall k \in A_j, \forall k' \in A_j, \forall j \in J, \)
\[
p_j^l = \frac{\pi_k}{\sum_{k'' \in A_j} \pi_{k''}} = \frac{\sum_{k'' \in A_j} \pi_{k''} / \pi_k}{\sum_{k'' \in A_j} \pi_{k''} / \pi_{k'}} = \frac{\sum_{k'' \in A_j} \pi_{k''} / \pi_k}{\pi_{k'}} = \frac{\pi_j^l}{\pi_{k'}^l} = \pi_k^l.
\]

(a’) ⇔ (b): \( \forall k \in A_j, \forall k' \in A_j, \forall j \in J, \)
\[
\beta_{kk'} = \frac{\pi_k}{\pi_{k'}} = \frac{\sum_{k'' \in A_j} \pi_{k''}}{\sum_{k'' \in A_j} \pi_{k''}} = \frac{\sum_{k'' \in A_j} \pi_j^l / \pi_{k''}^l}{\pi_{k'}^l} = \frac{\beta_{k'k}}{\pi_{k'}^l}.
\]

(b) ⇒ (c): \( \forall j \in J \) and for any function \( f(X) \),
\[
E[f(X)|A_j] = \sum_{k \in A_j} f(k)p_k^l = \sum_{k \in A_j} f(k)\pi_k^l = E_j[f(X)].
\]

(b) ⇐ (c): Let \( f(X) = I(X = k) \), where \( I(\cdot) \) is an indicator function and \( k \in A_j \). Then for \( \forall k \in A_j, \forall j \in J, \)
\[
p_k^l = E[I(X = k)|A_j] = E_j[I(X = k)] = \pi_k^l.
\]

\[\square\]

Condition (c) in Proposition 1 shows that under conservation of distribution, the change of probability measure does not affect the expected values, or in other words, a conditional expectation given the MC is in a specific subchain can be replaced by an expectation evaluated on a decomposed subchain. This is an essential property when implementing the total expectation theorem in MC settings.

For notational convenience, under conservation of distribution, we define \( \pi^l_{(k)} \), where a state \( k \) does not necessarily belong to a subchain \( j \):
\[
\pi^l_{(k)} = \beta_{kk'} \cdot \pi^l_{k'}, \forall k \in J, \forall k' \in A_j.
\]

Note that \( \pi^l_{(k)} \) does not depend on \( k' \) in subchain \( j \). Note also that \( \pi^l_j = \pi^l_k \) if \( k \in A_j \). Using this notation, conditions (a’) and (b’) can be re-represented as follows:

(a’’)
\[
\beta_{kk'} = \frac{\pi^l_{(k)}}{\pi^l_{(k')}} \cdot \pi^l_{k'}, \forall k \in S, \forall k' \in S, \forall j \in J.
\]

(b’’)
\[
P_j = \frac{\pi_k}{\pi^l_{(k)}} \cdot \pi^l_{k'}, \forall k \in S, \forall j \in J.
\]
2.2. Total Expectation Theorem for a MC

When we adopt a divide-and-conquer approach to a MC, we want a decomposed subchain to faithfully represent a part of the full MC; i.e., we require that a stationary distribution is conserved after decomposition: $\pi_k \pi^j_k, \forall k \in A_j, \forall j \in J$. This conservation of distribution condition leads to the following theorem with help from Proposition[1]. The theorem is essentially the total expectation theorem in MC settings.

**THEOREM 1. (Total expectation theorem for a MC)**

Under the conservation of distribution condition, an expectation of the full MC can be represented by expectations and steady-state probabilities of subchains, where a reference state $k$ can be any state in the full MC:

(a) $E[f(X)] = \sum_{j \in J^+} \frac{E_j[f(X)]}{\pi^j_k} - \sum_{j \in J^-} \frac{E_j[f(X)]}{\pi^j_k}$,

or equivalently,

(b) $\frac{E[f(X)]}{\pi_k} = \sum_{j \in J^+} \frac{E_j[f(X)]}{\pi^j_k} - \sum_{j \in J^-} \frac{E_j[f(X)]}{\pi^j_k}$.

**Proof of Theorem**[2]. We prove the theorem in two parts. The first part shows the equivalence of the two expressions. The second part derives expression (b).

(first part)

(a) $\Leftrightarrow$ (b): By letting $f(X) = 1$ in (b), we obtain the identity

$\frac{1}{\pi_k} = \sum_{j \in J^+} \frac{1}{\pi^j_k} - \sum_{j \in J^-} \frac{1}{\pi^j_k}$.

Combining this with (b), we obtain (a).

(a) $\Rightarrow$ (b): By letting $f(X) = I(X = k)$, we have the following identities:

$E[f(X)] = E[I(X = k)] = \pi_k$

and

$E_j[f(X)] = E_j[I(X = k)] = \pi^j_k I(k \in A_j)$.

Hence, (a) is reduced to the following:

$\pi_k = \frac{\sum_{j \in J^+} I(k \in A_j) - \sum_{j \in J^-} I(k \in A_j)}{\sum_{j \in J^+} \frac{1}{\pi^j_k} - \sum_{j \in J^-} \frac{1}{\pi^j_k}} = \frac{1}{\sum_{j \in J^+} \frac{1}{\pi^j_k} - \sum_{j \in J^-} \frac{1}{\pi^j_k}}$. 
where we have used the definitions of $J^+$ and $J^-$ (Equation (1)). Combining this result with (a), we obtain (b).

(Second part) Next, we prove a variant of expression (b), whose right and left hand sides are multiplied by $\pi_k$. Using Proposition 1 and the definitions of $p_k^j$, $E[\cdot|A_j]$, $J^+$, and $J^-$, we can show that the right hand side of the variant (RHS) matches with the left hand side of the variant (LHS):

$$RHS = \sum_{j \in J^+} \frac{E[f(X)|A_j]}{\pi^j_{(k)}} \cdot \pi_k - \sum_{j \in J^-} \frac{E[f(X)|A_j]}{\pi^j_{(k)}} \cdot \pi_k$$

$$= \sum_{j \in J^+} E[f(X)|A_j]P_j - \sum_{j \in J^-} E[f(X)|A_j]P_j$$

$$= \sum_{j \in J^+} \sum_{k \in A_j} f(k)p_k^j P_j - \sum_{j \in J^-} \sum_{k \in A_j} f(k)p_k^j P_j$$

$$= \sum_{j \in J^+} \sum_{k \in A_j} f(k)\pi_k - \sum_{j \in J^-} \sum_{k \in A_j} f(k)\pi_k$$

$$= \sum_{j \in J^+} \sum_{k \in S} I(k \in A_j)f(k)\pi_k - \sum_{j \in J^-} \sum_{k \in S} I(k \in A_j)f(k)\pi_k$$

$$= \sum_{k \in S} \left( \sum_{j \in J^+} I(k \in A_j) - \sum_{j \in J^-} I(k \in A_j) \right) f(k)\pi_k$$

$$= \sum_{k \in S} f(k)\pi_k = E[f(X)] = LHS.$$  

$\square$

**Remark 1.** Theorem 1 includes a reference state $k$, which can be arbitrarily chosen from the set $S$ but must be fixed throughout the calculation. This reference state can be replaced by a reference “set” $K$, which includes multiple states. The extension from a reference state $k$ to a reference set $K$ in Theorem 1 is straightforward by observing that $\pi_k$ and $\pi_k^j$ in Proposition 1 can be replaced by $\pi_K(\triangleq \sum_{k \in K} \pi_k)$ and $\pi_K^j(\triangleq \sum_{k \in K \subseteq A_j} \pi_k^j)$, respectively. $\square$

Theorem 1 implies that a function $\mu_k(A_j) = E_j[f(X)]/\pi_k^j$ satisfies an additivity property for a disjoint set of subchains. Specially, by considering the $\sigma$-algebra $F$ generated by a set $\{A_j, j \in J^+\}$, the function $\mu_k(\cdot)$ satisfies the following proposition:

**Proposition 2.** (Additivity property)

Assume that conservation of distribution holds after a decomposition procedure. Let $\{A_j, j \geq 1\}$ be a collection of disjoint subchains (i.e., $J^+ = \{j | j \geq 1\}$ and $J^- = \emptyset$) in the $\sigma$-algebra $F$. Then a function $\mu_k(A_j) = E_j[f(X)]/\pi_k^j$ satisfies the countable additivity property:

$$\mu_k\left( \bigcup_{j=1}^{\infty} A_j \right) = \sum_{j=1}^{\infty} \mu_k(A_j).$$
Proposition 2 is immediately obtained from Theorem 1. This countable additivity property, together with assumptions, \( \mu_k(\emptyset) = 0 \) and \( \mu_k(A) \geq 0, \forall A \in \mathcal{F} \), indicates that the function \( \mu_k(\cdot) \) satisfies the definition of a measure (see, for example, [Halmos 1950] and [Grimmett and Stirzaker 2001]). Hence, \( \mu_k(\cdot) \) inherits the following important properties that any measure has:

**Proposition 3.** Let \( A_j \) be a subchain in the \( \sigma \)-algebra \( \mathcal{F} \). The measure \( \mu_k(\cdot) \) satisfies the following properties:

1. For any subchain \( A_j \) and its complement \( A_j^c = S \setminus A_j \),
   \[
   \mu_k(S) = \mu_k(A_j) + \mu_k(A_j^c).
   \]

2. If subchains \( A_2 \supseteq A_1 \), then
   \[
   \mu_k(A_2) = \mu_k(A_1) + \mu_k(A_2 \setminus A_1) \geq \mu_k(A_1).
   \]

3. For any two subchains \( A_1 \) and \( A_2 \), \( \mu_k(A_1 \cup A_2) = \mu_k(A_1) + \mu_k(A_2) - \mu_k(A_1 \cap A_2) \). In particular, if both subchains are disjoint, then \( \mu_k(A_1 \cup A_2) = \mu_k(A_1) + \mu_k(A_2) \).

**Proof of Proposition 3.** (1) Since \( A \cup A^c = S \) and \( A \cap A^c = \emptyset \), Theorem 1 says \( \mu_k(S) = \mu_k(A \cup A^c) = \mu_k(A) + \mu_k(A^c) \). (2) Denote the full MC by \( A_2 \). Considering the two disjoint sets \( A_1 \) and \( A_2 \setminus A_1 \) in Theorem 1, we have \( \mu_k(A_2) = \mu_k(A_1) + \mu_k(A_2 \setminus A_1) \). (3) Again, considering the two disjoint sets \( A_1 \) and \( A_2 \setminus A_1 \) and using (2) above, we have \( \mu_k(A_1 \cup A_2) = \mu_k(A_1) + \mu_k(A_2 \setminus A_1) = \mu_k(A_1) + \mu_k(A_2) - \mu_k(A_1 \cap A_2) \). □

Proposition 3 guarantees that any set of subchains that is convenient for our analysis can satisfy the additivity property: Subchains can be disjoint or overlapping, and their number can be finite or infinite. Specifically, we can consider the following three special cases:

1. A set of subchains that are all disjoint \( (A_i \cap A_j = \emptyset) \) for all pairs \((i, j)\) in \( J^+ \) satisfying \( i \neq j \):
   \[
   \mu_k(S) = \sum_{j \in J^+} \mu_k(A_j).
   \]

2. A set of subchains that are nested \( (A_i \subseteq A_{i+1}, \forall i \in J^+ = \{1, 2, \ldots\}) \):
   \[
   \mu_k(S) = \mu_k(A_1) + \sum_{i \geq 1} \mu_k(A_{i+1} \setminus A_i).
   \]

3. A set of subchains in which states are elements of at most two subchains \((A_i \cap A_j \neq \emptyset)\) for some pairs \((i, j)\) in \( J^+ \) satisfying \( i \neq j \), but \((A_i \cap A_j) \cap (A_{i'} \cap A_{j'}) = \emptyset \) for all pairs \((i, j)\) and \((i', j')\) in \( J^+ \) satisfying \( i \neq j, i' \neq j', \) and \((i, j) \neq (i', j')\):
   \[
   \mu_k(S) = \sum_{j \in J^+} \mu_k(A_j) - \sum_{i, j \in J^+, i < j} \mu_k(A_i \cap A_j).
   \]
When we decompose a MC, a set of decomposed subchains often falls into one of these three cases. But before discussing examples of such subchains, we present a scheme to satisfy conservation of distribution when decomposing the full MC. This conservation of distribution is required for Theorem 1 to hold.

3. Termination

If we arbitrarily decompose a MC into subchains, each subchain independently analyzed is likely to have a different stationary distribution from its portion of the stationary distribution of the full MC: Proposition 1 may not hold. This is because a subchain extracted from the full MC loses transitions between states in the subchain and states outside of the subchain. Such a loss can affect the distribution of a decomposed subchain; specifically, it could violate the conservation of distribution property. Therefore, in order to conserve the stationary distribution of a decomposed subchain, we must alter transitions at states that lose transitions. We call our alteration procedure termination.

The simplest example of termination is truncation, in which no alteration needs to be made for states that lose transitions. Truncation does not always conserve the stationary distribution of a decomposed subchain, unless the MC has a special structure (e.g., quasi-reversibility). In contrast, termination, if appropriately chosen, can always conserve a stationary distribution. Note that, for obvious reason, we require termination to conserve ergodicity of a subchain as well (i.e., a terminated subchain should have a single communicating class).

In this section, we demonstrate one procedure to obtain a termination scheme that satisfies both conservation of distribution and ergodicity. (There always exist infinitely many possible termination schemes. But it is not necessary to find all such schemes; to make the total expectation theorem work, we need only one termination scheme.)

3.1. Necessary and Sufficient Condition for a Decomposed Subchain to Conserve a Stationary Distribution

We continue our analysis of the set \( \{ A_j, \ j \in J \} \), \( J = J^+ \cup J^- \), which satisfies \( S = \bigcup_{j \in J^+} A_j \).

We denote a transition rate from state \( k \) to state \( k' \) in the full MC as \( q_{k,k'} \). Let the set of states in \( A_j \) that lose transitions due to decomposition be denoted as the boundary set: \( \partial A_j = \{ k \in A_j | q_{k,k'} + q_{k',k} > 0, \exists k' \in A_j^c \} \). Let the set of states in \( A_j \) that is not in \( \partial A_j \) be denoted as the interior set: \( \text{int}(A_j)(= A_j \setminus \partial A_j = \{ k \in A_j | q_{k,k'} + q_{k',k} = 0, \forall k' \in A_j^c \}) \). Hence, \( \partial A_j \cup \text{int}(A_j) = A_j \) and \( \partial A_j \cap \text{int}(A_j) = \emptyset \). (Note that \( \partial A_j \neq \emptyset \) since a subset \( A_j(\subset S) \) should communicate with its complement \( A_j^c(= S \setminus A_j) \) under our ergodicity assumption for the full MC.)
We consider terminating boundary states of a decomposed subchain: Adding new transitions $\Delta q_{k,k'}^j$ for some $k,k' \in \partial A_j$ in subchain $j$. (If $\Delta q_{k,k'}^j = 0, \forall k \in \partial A_j, \forall k' \in \partial A_j$, then we call such a termination scheme truncation.) Termination of a decomposed subchain $j$ is added only to the states in $\partial A_j$ and not to any states in $\text{int}(A_j)$.

Under a termination scheme, we claim that the conservation of distribution condition shown in Proposition 1 is equivalent to the boundary condition, as described in the following proposition:

**Proposition 4. (Boundary Condition)**

Under a termination scheme, conservation of distribution (Proposition 1) holds if and only if any of the following conditions are satisfied:

(a) $\pi_k \propto \pi_k^j, \forall k \in \partial A_j, \forall j \in J.$

(b) $\beta_{kk'} = \frac{\pi_k^j}{\pi_{k'}^j}, \forall k \in \partial A_j, \forall k' \in \partial A_j, \forall j \in J.$

(c) $\beta_{kk'} = \frac{\pi_k^j}{\pi_{k'}^j}, \forall k \in \partial A_j, \forall j \in A_j.$

(d) $p_k^j = \pi_k^j, \forall k \in \partial A_j, \forall j \in J.$

(e) $P_j = \frac{\pi_k^j}{\pi_{k}^j}, \forall k \in \partial A_j, \forall j \in J.$

**Proof of Proposition 4.**

The equivalence of all conditions in this proposition can be proved in the same way as the proof of Proposition 1. Here, we only prove the equivalence between conditions (a) in Proposition 1 and (d) in Proposition 4.

(a) $\Leftrightarrow$ (d): Let a steady-state distribution of the full MC be a row vector $\pi = (\pi_{\text{int}(A_j)}, \pi_{\partial A_j}, \pi_{A_j^c})$, which is divided into three row vectors according to where each state lies. Note that $\pi$ is uniquely determined due to the requirement of ergodicity of the full MC. Let a corresponding transition matrix be $P = (P_{ji}), i,j \in \{\text{int}(A_j), \partial A_j, A_j^c\}$. Since $\pi$ satisfies the equation $\pi = \pi P$, flow balance equations for the set $\text{int}(A_j)$ should follow the equation $\pi_{\text{int}(A_j)} = \pi_{\text{int}(A_j)} P_{\text{int}(A_j),\text{int}(A_j)} + \pi_{\partial A_j} P_{\partial A_j,\text{int}(A_j)}$. The distribution $\pi_{\text{int}(A_j)}$ should be uniquely solvable in terms of $\pi_{\partial A_j}$. (If not, there exist multiple solutions for $\pi$.) Therefore, $\pi_{\text{int}(A_j)} = \pi_{\partial A_j} P_{\partial A_j,\text{int}(A_j)} (I - P_{\text{int}(A_j),\text{int}(A_j)})^{-1}$.

Notice that both $P_{\partial A_j,\text{int}(A_j)}$ and $P_{\text{int}(A_j),\text{int}(A_j)}$ are not altered by termination (nor by truncation). Hence, the same equation holds for both the full MC and a decomposed subchain $j$ with termination. Since (d) says $\pi_k \propto \pi_k^j, \forall k \in \partial A_j$, we know $\pi_k \propto \pi_k^j, \forall k \in A_j$.

(a) $\Rightarrow$ (d): This is obvious because $\partial A_j \subset A_j$. \(\square\)
According to Proposition 4, the condition for conservation of distribution reduces to the boundary condition:

$$\pi_k \propto \pi_k^j, \forall k \in A_j \Leftrightarrow \pi_k \propto \pi_k^j, \forall k \in \partial A_j.$$ 

This is intuitively obvious because termination (or truncation as a special case) only changes flow balance equations at boundary states. As long as the boundary distribution is maintained, the entire distribution of a decomposed subchain is conserved. In this paper, we use the terms “the conservation of distribution condition” and “the boundary condition” interchangeably as they are equivalent.

How can we satisfy the boundary condition? Needless to say, the simplest way to satisfy this condition is to impose $\pi_k^j$ follow $\pi_k \propto \pi_k^j, \forall k \in \partial A_j$, if we know how $\pi_k$ is distributed at the boundary states. This is often possible when these boundary states also belong to another subchain, for which we know the distribution. However, in general, we need to find $\pi_k^j$ at boundary states by solving an independent subchain $j$ with termination. We next discuss the condition for termination that ensures the boundary condition.

3.2. Partial Flow Conservation: An Equivalent Condition to the Boundary Condition

One obvious conclusion we can draw from Proposition 4 is that if the number of states in a boundary set is one for a subchain (i.e., $|\partial A_j| = 1, j \in J$), then the boundary condition for the subchain is automatically satisfied, and therefore, the stationary distribution of this subchain is conserved regardless of any termination schemes we apply, including the simplest termination scheme: Truncation. Intuitively, if a subchain connects to the rest of the chain through a single state, the flow out from the subchain through the single state must be balanced by the flow into the subchain through the same single state, in which case a loss of flows due to decomposition does not alter the distribution of a decomposed subchain.

However, if there is more than one boundary state we need to control flows at each boundary state to potentially compensate for a loss of transitions due to decomposition. Specifically, we need to conserve the net outflow from each boundary state in order to conserve a boundary distribution after decomposition. We call such a condition partial flow conservation, and claim that termination must satisfy the partial flow conservation condition:

**Lemma 1.** *(Partial flow conservation)*

The stationary distribution is conserved after decomposition if and only if termination conserves the partial flow at boundary states:
\[ (f) \quad \pi_k \sum_{k' \in A_j^c} q_{k,k'} - \sum_{k' \in A_j^c} \pi_{k'} q_{k,k'} = \pi_k \sum_{k' \in \partial A_j} \Delta q_{k,k'}^{j} - \sum_{k' \in \partial A_j} \pi_{k'} \Delta q_{k,k'}^{j}, \forall k \in \partial A_j, \forall j \in J. \]

To prove Lemma 1, we use the following proposition: It provides a set of equations that holds at boundary states of the full MC and a terminated subchain. As the proof is straightforward, we omit it.

**Proposition 5.** Steady-state probabilities at boundary states for the full MC and a terminated subchain satisfy the following equations:

1) Steady-state equations at boundary states for the full MC:

\[ \pi_k \left( \sum_{k' \in A_j} q_{k,k'} + \sum_{k' \in \partial A_j} \Delta q_{k,k'}^{j} \right) \overset{\text{d}}{=} \sum_{k' \in A_j} \pi_{k'} q_{k',k} + \sum_{k' \in \partial A_j} \pi_{k'} \Delta q_{k',k}^{j}, \forall k \in \partial A_j. \quad (2) \]

2) Steady-state equations at boundary states for a terminated subchain:

\[ \pi_k^j \left( \sum_{k' \in A_j} q_{k,k'} + \sum_{k' \in \partial A_j} \Delta q_{k,k'}^{j} \right) \overset{\text{d}}{=} \sum_{k' \in A_j} \pi_{k'} q_{k',k} + \sum_{k' \in \partial A_j} \pi_{k'} \Delta q_{k',k}^{j}, \forall k \in \partial A_j. \quad (3) \]

**Proof of Lemma 1.** Note that Equations (2) and (3) in Proposition 5 always hold. We show the equivalence of conditions (d) in Proposition 4 and (f) in Lemma 1 via Proposition 5.

\( (d) \Rightarrow (f): \) If (d) holds, then a set of probabilities \( \pi_k, \forall k \in \partial A_j \), should satisfy not only Equation (2) but also Equation (3):

\[ \pi_k \left( \sum_{k' \in A_j} q_{k,k'} + \sum_{k' \in \partial A_j} \Delta q_{k,k'}^{j} \right) = \sum_{k' \in A_j} \pi_{k'} q_{k',k} + \sum_{k' \in \partial A_j} \pi_{k'} \Delta q_{k',k}^{j}, \forall k \in \partial A_j, \]

which is equivalent to Equation (3) for \( \pi_k^j \). Since the steady-state equations at boundary states for both the full MC and a terminated subchain are equivalent, under the ergodicity assumption for a terminated subchain, we should obtain the same solutions at boundary states (up to a normalization constant). Hence, (d) holds.

\( (d) \Rightarrow (f): \) If (d) holds, then a set of probabilities \( \pi_k, \forall k \in \partial A_j \), should satisfy not only Equation (2) but also Equation (3):

\[ \pi_k \left( \sum_{k' \in A_j} q_{k,k'} + \sum_{k' \in \partial A_j} \Delta q_{k,k'}^{j} \right) = \sum_{k' \in A_j} \pi_{k'} q_{k',k} + \sum_{k' \in \partial A_j} \pi_{k'} \Delta q_{k',k}^{j}, \forall k \in \partial A_j. \]

Combining this equation with Equation (2) for \( \pi_k \) and eliminating terms with summation over a set \( A_j \), we obtain (f). \( \Box \)
Lemma 1 says that if added transitions (termination) $\Delta q_{k,k'}$ can replicate the net outflow from a subchain in the full MC, then the steady-state distribution of the terminated subchain will conserve the stationary distribution of the full MC. Note that if termination (or truncation as a special case) is chosen to satisfy partial flow balance conservation (Lemma 1), then the termination automatically satisfies global flow balance as well. This is expected because termination does not add transitions between a subchain $j$ and its complement. This leads to the following corollary:

**Corollary 1.** *Global flow balance holds if termination satisfies partial flow conservation:*

$$
\sum_{k \in \partial A_j} \left( \pi_k \sum_{k' \in A_j^c} q_{k,k'} - \sum_{k' \in A_j^c} \pi_{k'} q_{k',k} \right) = 0.
$$

**Proof of Corollary 1**

Using Lemma 1, we obtain:

$$
\sum_{k \in \partial A_j} \left( \pi_k \sum_{k' \in A_j^c} q_{k,k'} - \sum_{k' \in A_j^c} \pi_{k'} q_{k',k} \right) = \sum_{k \in \partial A_j} \left( \pi_k \sum_{k' \in \partial A_j} \Delta q_{k,k'} - \sum_{k' \in \partial A_j} \pi_{k'} \Delta q_{k',k} \right) \\
= \sum_{k,k' \in \partial A_j} \pi_k \Delta q_{k,k'} - \sum_{k,k' \in \partial A_j} \pi_{k'} \Delta q_{k',k} \\
= 0.
$$

□

A special case of partial flow conservation (Lemma 1) is partial flow balance, which was first introduced by Whittle (1968) and was discussed extensively in the context of queueing networks by Kelly (1979):

$$
\pi_k \sum_{k' \in A_j^c} q_{k,k'} - \sum_{k' \in A_j^c} \pi_{k'} q_{k',k} = 0, \forall k \in \partial A_j.
$$

When this partial flow balance condition holds, we can satisfy Lemma 1 by selecting $\Delta q_{k,k'} = 0, \forall (k, k') \in \partial A_j$, which means that partial flow conservation, the boundary condition, and conservation of distribution are all satisfied by a simple truncation scheme. This property is known as the *state truncation property* (see Nelson 1995). (Note: If ergodicity is not maintained by truncation, we still need termination.) However, in general, partial flow balance does not always hold and/or ergodicity may not be maintained by truncation. In such cases, a non-trivial termination scheme that satisfies the partial flow conservation condition is required.

### 3.3. Termination That Satisfies Partial Flow Conservation

In this subsection, we discuss a scheme to obtain termination that satisfies partial flow conservation (Lemma 1). We first show one possible termination scheme that works for any general MC. We then apply the scheme to more specific cases to obtain simplified expressions.
From Lemma [1], the simplest condition that we can require for termination to hold is the following condition. (Note that we continue to require termination to maintain ergodicity of a decomposed subchain.)

**Proposition 6.** *(Termination condition sufficient to satisfy Lemma [1]):*

The boundary distribution is conserved if termination $\Delta q^i_{k,k'}$ satisfies the following flow conservation conditions at all boundary states:

1. **Outflow (from a subchain) condition:**
   \[
   \sum_{k' \in A^c} q_{k,k'} = \sum_{k' \in \partial A_j} \Delta q^i_{k,k'}, \forall k \in \partial A_j.
   \]

2. **Inflow (to a subchain) condition:**
   \[
   \sum_{k' \in A^c} \pi_{k'} q_{k',k} = \sum_{k' \in \partial A_j} \pi_{k'} \Delta q^i_{k,k'}, \forall k \in \partial A_j.
   \]

**Proof of Proposition 6.** By inspection, the termination scheme satisfying these two flow conditions satisfies Lemma [1]. Therefore, the boundary distribution and hence the stationary distribution of a decomposed subchain must be conserved. □

**Remark 2.** Termination that satisfies Proposition [6] is not the only possible termination scheme that satisfies Lemma [1]. In fact, there are countless possible variations of termination that satisfy Lemma [1] but not Proposition [6]. For example, we know we can arbitrarily add or drop self-transitions from termination without affecting Lemma [1]. However, such a change in termination is not allowed by Proposition [6]. Another example is that if we could find another termination $\Delta q^j_{k,k'}$ that satisfies

\[
\pi_k \sum_{k' \in \partial A_j} \Delta q^j_{k,k'} = \sum_{k' \in \partial A_j} \pi_{k'} \Delta q^j_{k,k'}, \forall k \in \partial A_j,
\]

then the new termination $\Delta q^i_{k,k'} + \Delta q^j_{k,k'}$ does not satisfy Proposition [6] but satisfies Lemma [1]. □

We can always find at least one possible termination scheme that satisfies Proposition [6] and the ergodicity requirement. This termination scheme can be simplified for two special cases: (1) when there exists only one state at the boundary that has inflows from outside; (2) when there exists only one state in the external set ($A^c_j$). We can summarize the result in the following corollary:

**Corollary 2.** *(Termination scheme)*

The following termination $\Delta q^i_{k,k'}$ satisfies Proposition [6]

\[
\Delta q^i_{k,k'} = \frac{\sum_{m \in A^c_j} \pi_m q_{m,k'}}{\sum_{k'' \in \partial A_j} \sum_{m \in A^c_j} \pi_m q_{m,k''}} \cdot \sum_{m \in A^c_j} q_{k,m}, \forall k,k' \in \partial A_j.
\]  (4)
π_m, m ∈ A^c_j, can be dropped in the following two special cases:

(1) If there is only one state s ∈ ∂A_j such that \( \sum_{m \in A^c_j} \pi_m q_{m,s} > 0 \) and for any other states k(≠ s) ∈ ∂A_j, \( \sum_{m \in A^c_j} \pi_m q_{m,k} = 0 \) holds. Then the termination scheme (Equation (4)) is reduced to:

\[
\Delta q_{j,s} = \sum_{m \in A^c_j} q_{k,m}, \forall k \in \partial A_j,
\]

\[
\Delta q_{j,k'} = 0, \forall k \in \partial A_j, \forall k'(≠ s) \in \partial A_j.
\]

(2) If there is only one state m ∈ A^c_j, then the termination scheme (Equation (4)) is reduced to:

\[
\Delta q_{j,k,k'} = \frac{q_{m,k'}}{\sum_{k'' \in \partial A_j} q_{m,k''}} \cdot q_{k,m}, \forall k, k' \in \partial A_j.
\]

Proof of Corollary 2. We can confirm that Equation (4) satisfies Proposition 6. Since the added transitions (termination) connect all boundary states, this termination scheme fulfills the ergodicity requirement of a decomposed subchain. It is straightforward to derive the two special cases from Equation (4). □

Remark 3. Note again that the termination we derive in Corollary 2 can be modified to include or exclude self-transitions because such a modification does not affect Lemma 1. □

Remark 4. If we can partition the external states (states in A^c_j) into multiple non-communicating classes, Corollary 2 can be applied to each class, and we can obtain termination schemes for multiple classes. Aggregation of these terminations represents the correct termination that satisfies Lemma 1. □

The two special cases in Corollary 2 are especially useful when we want to derive analytical solutions, because closed-form solutions are easy to obtain when appropriate terminations are determined without the knowledge of the distribution of external states. The first special MC structure of the two cases is well-known and called a single-input superstate decomposable Markov chain (SISDMC), which was first discussed by Feinberg and Chiu (1987). However, in general, we may have both multiple input states and multiple external states, and hence, we need to resort to a recursive method to derive an appropriate termination. To make Corollary 2 suitable for use in a numerical algorithm, we rewrite Equation (4) using distributions of decomposed subchains (\( \pi^i_k \)) rather than a distribution of the full MC (\( \pi_k \)). Let the probability that a state is in subchain \( j \) be \( w_j \). We obtain the following proposition that can be used for a numerical algorithm:

Proposition 7. If decomposed subchains (\( A_j, j ∈ J^+ \)) are disjoint, the following set of termination \( \Delta q^j_{k,k'} \) satisfies Proposition 6:

\[
\Delta q^j_{k,k'} = \frac{\sum_{i(≠ j)} w_i \sum_{m \in A^i_i} \pi^i_m q_{m,k'}}{\sum_{k'' \in \partial A_j} \sum_{i(≠ j)} w_i \sum_{m \in A^i_i} \pi^i_m q_{m,k''}} \cdot \sum_{i(≠ j)} \sum_{m \in A^i_i} q_{k,m}, \forall k, k' \in \partial A_j, \forall j \in J^+,
\]
where \( w_j \) should satisfy global flow balance condition (Corollary 7):

\[
\sum_{k \in \partial A_j} \sum_{i \neq j} w_i \sum_{m \in A_i} \pi_{m,i} \sum_{k \in \partial A_j} q_{m,k}, \forall j \in J^+,
\]

which is equivalent to

\[
w_j = \frac{\sum_{i \neq j} w_i \sum_{m \in A_i} \pi_{m,i} \sum_{k \in \partial A_j} q_{m,k}}{\sum_{k \in \partial A_j} \sum_{i \neq j} \sum_{m \in A_i} \pi_{m,i} \sum_{k \in \partial A_j} q_{m,k}}, \forall j \in J^+.
\]

**Proof of Proposition** Using Proposition 1, \( \pi_{m,m} \in A^c_j \), in Equation (4) can be replaced by \( w_i \pi_{m,i} \in A^c_j \cap A_i, i \neq j \). Hence, if subchains are disjoint, we can re-write Equation (4) as Proposition 7. \( \square \)

### 4. Applications of Our MC Decomposition Method

In this section, we demonstrate the benefits of our method by applying it to several MCs representing queueing systems. We first analyze simple MCs using a set of overlapping or nested subchains to show relationships among performance indicators. Next we analyze a more complex MC, which represents a queueing system with adjustable staffing levels. This queueing system is used to analyze the performance of border gates/toll booths (with no customer abandonment). The MCs that represent these systems are complicated and thus exact solutions or accurate approximations are often difficult to obtain (see Zhang 2009 and Bhandari et al. 2008). Our method simplifies the process of problem solving, making it possible to derive exact solutions by summing up the solutions of decomposed subchains with appropriate terminations.

#### 4.1. Performance Indicators and Their Relationships

We derive relationships among performance indicators, such as the blocking probability \( P_{\text{block}} \) (the steady-state probability of the end state of the chain), the queueing probability \( P_Q \), and the number of customers in a queue \( L_Q \). Such relationships are often difficult to establish, but Theorem 1 enables us to find those relationships without explicitly deriving the performance indicators.

We consider two cases: (1) Subchains share a single state with each other (\(|A_i \cap A_j| = 1 \) for \( i \neq j \)) and (2) Subchains are nested (\( A_i \cap A_j = A_i \) or \( A_j \) for \( i \neq j \)).

**4.1.1. Subchains Sharing a Single State** Consider decomposing a general MC into subchains 1 and 2, which share a single state \( s \). Define \( A_3 = A_1 \cap A_2 = \{s\} \). Let \( J^+ = \{1, 2\} \) and \( J^- = \{3\} \). As long as proper termination is applied to all subchains, using Theorem 1 we obtain:

\[
\frac{E[f(X)]}{\pi_s} = \frac{E_1[f(X)]}{\pi_{s1}} + \frac{E_2[f(X)]}{\pi_{s2}} - f(s), \text{ for any } f(X),
\]

(8)
where $\pi_s^3 = 1$ is used above. This is a general property that always holds regardless of what MC we deal with, how we decompose it, and which function of a MC state we use.

To derive some practically useful properties, consider a simple queueing system which has a chain corresponding to no queue (subchain 1) and a chain corresponding to queue (subchain 2). A state $s$ could be a transitional state shared by both subchains 1 and 2, representing a situation where there is no queue but any new arrival will be put in a queue. (As an example, we can split an M/M/s queue into an M/M/s/s queue and an M/M/1 queue where a state $s$ is shared in the middle.) Denote the queueing probability of this system (i.e., the probability that a new arrival is put in a queue) as $P_Q$, the average number of people in the queue as $L_Q$, and the average number of people in the queue evaluated in subchain 2 as $L_Q^2$. Also, let an operator $N_Q$ represent the number of people in the queue. By applying $f(X) = 1$, $I(k \in A_2)$, and $N_Q$ to Equation (8), we obtain, correspondingly,

$$\frac{1}{\pi_s} = \frac{1}{\pi_s} + \frac{1}{\pi_s^2} - 1, \quad \frac{P_Q}{\pi_s} = \frac{1}{\pi_s}, \quad \text{and} \quad \frac{L_Q}{\pi_s} = \frac{L_Q^2}{\pi_s}. \quad (9)$$

Relations among performance indicators are directly obtained from Equation (9). We denote $\pi_s^1$ as blocking probability $P_{\text{block}}$, which represents the probability that a new arrival is blocked to enter a system represented by decomposed subchain 1. By eliminating $\pi_s$, we obtain

$$P_{\text{block}} = \frac{\pi_s^2 P_Q}{1 - (1 - \pi_s^2) P_Q} = \frac{L_Q}{\frac{L_Q}{\pi_s^2} - \left(\frac{1}{\pi_s^2} - 1\right) L_Q}. \quad (10)$$

Equation (10) holds for any general MC that can be split into no queue and queue subchains, which share a single transitional state $s$. To find the relationship among $P_{\text{block}}$, $P_Q$, and $L_Q$, we do not need to derive their actual representations; instead, we need concrete representations of $\pi_s^2$ and $L_Q^2$ in Equation (10). In particular,

1) If subchain 2 is an M/M/1 queue with $\pi_s^2 = 1 - \rho$ where $\rho < 1$, then by using the property $L_Q^2 = \rho/(1 - \rho)$, we obtain

$$P_{\text{block}} = \frac{(1 - \rho) P_Q}{1 - \rho P_Q} = \frac{(1 - \rho)^2}{\rho} \frac{L_Q}{1 - (1 - \rho) L_Q}. \quad (11)$$

This relationship is known to hold for Erlang B/C models when subchain 1 is an M/M/s/s queue, and appears in many textbooks (see, for example, [Harchol-Balter, 2013]). However, we proved that this relationship holds for any subchain 1, not just when subchain 1 is an M/M/s/s queue.

2) If subchain 2 is an M/M/1/k queue with $\pi_s^2 = 1/\sum_{n=0}^{k} \rho^n$, then by using the property $L_Q^2 = \sum_{n=0}^{k} n\rho^n \pi_s^2$, we obtain

$$P_{\text{block}} = \frac{(1 - \rho) P_Q}{1 - \rho P_Q - \rho^{k+1}(1 - P_Q)} = \frac{(1 - \rho)^2}{\rho} \frac{L_Q}{1 - \rho^{k}(1 + k(1 - \rho)) - (1 - \rho)(1 - \rho^k) L_Q}, \quad (12)$$
where we have used the following well-known formulae:
\[
\sum_{n=0}^{k} \rho^n = \frac{1 - \rho^{k+1}}{1 - \rho} \quad \text{and} \quad \sum_{n=0}^{k} n \rho^n = \rho \frac{\partial}{\partial \rho} \left( \sum_{n=0}^{k} \rho^n \right) = \rho \frac{\partial}{\partial \rho} \left( \frac{1 - \rho^{k+1}}{1 - \rho} \right) = \rho \cdot \frac{1 - \rho^k (1 + k(1 - \rho))}{(1 - \rho)^2}.
\]

Equation (12) is a new relationship. As above, we proved that this relationship holds for any subchain 1, not just when subchain 1 is an M/M/s/s queue. We can easily confirm that Equation (12) converges to Equation (11) at the limit of \(k \to \infty\).

3) Let subchain 2 be a queueing system with discouraged arrivals, which assumes a harmonic discouragement of arrivals with respect to the number present in the system: \(\lambda_k = \alpha/(k+1), k = 0, 1, 2, \ldots\) and \(\mu_k = \mu, k = 1, 2, 3, \ldots\) for subchain 2. Let \(\rho = \alpha/\mu\). According to Kleinrock (1975), we know \(\pi_x^2 = e^{-\rho}\) and \(L_Q^2 = \rho\), from which we obtain
\[
P_{\text{block}} = \frac{e^{-\rho} P_Q}{1 - (1 - e^{-\rho}) P_Q} = \frac{L_Q}{\rho e^\rho - (e^\rho - 1) L_Q}.
\]
Equation (13) is a new relationship, which holds for any subchain 1.

4.1.2. Nested Subchains Consider a set of nested MCs \(\{A_k : k = 0, 1, 2, \cdots\}\). \(A_k\) is composed of \(k+1\) states: \(A_k = \{0, 1, 2, \ldots, k\}\), which satisfies \(A_k \supset A_{k-1}\) and \(\{k\} = A_k \setminus A_{k-1}\) for \(\forall k \in \mathbb{Z}^+\). Note that \(A_0 = \{0\}\). Assume that the boundary condition (Proposition 4) is always satisfied throughout the analysis. That is, every time we decompose a MC, we assume that an appropriate termination is applied to each decomposed subchain; hence, the steady-state distribution of the decomposed subchain is always proportional to the original steady-state before the decomposition is made. We are interested in finding a recursive equation for the steady-state probability of state \(k\) in \(A_k\). We decompose \(A_k\) into \(A_{k-1}\) and \(\{k\}\), and apply Theorem 1. Let state \(k\) be the reference state. For any function \(f(X)\) of states, the following recursive equation holds:
\[
\frac{E_k[f(X)]}{\pi_k^k} = \frac{E_{k-1}[f(X)]}{\beta_{k,k-1} \pi_{k-1}^{k-1}} + f(k), k \in \mathbb{Z}^+.
\]

Remark 5. By repeating the recursive process and utilizing the properties \(\beta_{k,i} \cdot \beta_{i,j} = \beta_{k,j}\) and \(\beta_{k,k} = 1\), Equation (14) is reduced to
\[
\frac{E_k[f(X)]}{\pi_k^k} = \sum_{i=0}^{k} \frac{f(i)}{\beta_{k,i}}.
\]
Equation (15) is immediately obtained by applying Theorem 1 to a set of decomposed “subchains” \(\{0\}, \{1\}, \cdots, \{k\}\), with reference state \(k\). Note also that Equation (15) is equivalent to the definition of expectation: By multiplying \(\pi_k^k\) to both sides of the equation, we recover
\[
E_k[f(X)] = \sum_{i=0}^{k} f(i) \pi_i^k.
\]

Equation (14) is a general property that holds for any finite nested MC and for any function of a MC state. To derive some practically useful properties, we consider a birth and death MC.
We do not need to derive their actual representations; instead, we need concrete representations of Equation (16) holds for any birth and death MC. To find recursive equations for each gate is operated by one staff member.) The most basic CBS policy uses two thresholds: The system is used to analyze the performance of border gates/toll booths/server farms. The staffing level is controllable, but operators may not want to adjust staffing levels frequently because changing staffing levels is costly. The simplest control policy for adjusting staffing levels is called the Congestion-Based Staffing (CBS) policy (Zhang 2009). Under this CBS policy, if the queue length is decreased, some toll gates are closed, and if increased, some gates are opened. (We assume that each gate is operated by one staff member.) The most basic CBS policy uses two thresholds: The...
lower threshold $n$, which switches the system to operate with a lower number of staff, and the upper threshold $N$, which switches the system to operate with a larger number of staff. Operators of toll gates seek the optimal $(n, N)$ combination to minimize the sum of three costs: The cost of staffing, customers’ waiting time, and switching staffing levels. The CBS policy is commonly used in practice, but its exact solution presented in Zhang (2009) is very complicated and hence an approximation is often sought. As an example to illustrate the simplicity of our MC decomposition method and its capability to cope with more extended models, we apply our method to derive simple, exact, analytical expressions for operational costs in the CBS model.

We assume there is no customer abandonment (neither reneging nor balking). The total number of toll gates are $c$, where $e$ out of $c$ ($c > e > 0$) are extra gates that can be opened or closed (all at the same time). Arrivals of vehicles to the toll gate occur at constant rate $\lambda$, according to a Poisson process. Service times at each gate are distributed exponentially with parameter $\mu$. Under the CBS policy, $e$ gates are opened when the total number of vehicles in the system reaches the upper threshold $N$ and are closed when the total number of vehicles in the system reaches the lower threshold $n$, where $N > n$ must hold. We limit our analysis to the case where $n \geq c$ and arrival rate is constant. Extensions of this model are discussed in Zhang (2009) (allowing the lower threshold $n$ to be less than $c$) and Bhandari et al. (2008) (time-varying arrival rate). The $n < c$ case in the CBS model can be handled using the same approach we present here. Analysis of the model with time-varying arrival rate requires a numerical decomposition approach, where we recursively identify a combination of $n$ and $N$ that optimizes the total operational cost. This analysis will be a potential research topic in the future.

**Figure 1** The CBS model.

Figure 1 shows the MC model corresponding to a toll gate system with a CBS policy. We call the upper part of the MC in Figure 1 chain A and the lower part chain B. Notice that chains A and B represent M/M/c-e/N-1 and M/M/c queues, and correspond to low capacity and high capacity...
modes, respectively. As Figure 1 shows, a system switches to high capacity mode (chain B) when an arrival occurs at state $N - 1$ in chain A (which we denote $N - 1_A$), and switches to low capacity mode when a departure of a vehicle occurs at state $n + 1$ in chain B ($n + 1_B$). A changeover cycle time corresponds to the time between two consecutive departure times from Chain A to Chain B. Define $c_g$, $c_w$, and $c_s$ as the cost of operation per gate per unit time, the waiting cost per vehicle per unit time, and a switching cost per cycle, respectively. Define $P_A$, $P_B (= 1 - P_A)$, $L_Q$, $f$, and $T_{cycle}$ as the probability to be in chain A, the probability to be in chain B, the expected number of vehicles in a queue, the frequency of changeover, and one changeover cycle time. The total cost is the sum of the following costs:

1. Average staffing cost:

$$C_{staff} = c_g \cdot (c - e + e \cdot P_B)$$

2. Average waiting cost:

$$C_{wait} = c_w \cdot L_Q$$

3. Average switching cost:

$$C_{switch} = c_s \cdot E[f] = c_s \cdot \frac{1}{E[T_{cycle}]} = c_s \cdot \lambda \cdot \pi_{N-1_A}$$

Note that $1/E[T_{cycle}] = \lambda \pi_{N-1_A}$ is used above. This is derived by applying Little’s law to the ergodic closed system with 1 job.

**4.2.1. Analysis of CBS Model** The total cost is a function of $\pi_{N-1_A}$, $P_B$ (or $P_A$), and $L_Q$. We obtain these quantities by utilizing our decomposition method. The first task is to decompose the full MC into five (partially overlapping) subchains: $A_1 = \{0_A, 1_A, \ldots, n - 1_A\}$, $A_2 = \{n_A, n + 1_A, \ldots, N - 1_A\}$, $A_3 = \{N - 1_A, n + 1_B\}$, $A_4 = \{n + 1_B, n + 2_B, \ldots, N_B\}$, $A_5 = \{N + 1_B, N + 2_B, \ldots\}$. We set $J^+ = \{1, 2, 3, 4, 5\}$ and $J^- = \{3\}$ to satisfy Equation (1). (Note: Subchain 3 can be dropped from both $J^+$ and $J^-$. However, we include subchain 3 since it is convenient to utilize subchain 3 in the analysis.) Let the reference state of the full MC be state $N - 1_A$. Let $N_Q$ be the operator for the number of vehicles in a queue. By plugging $f(X) = 1$, $I_B(= I(k \in A_4 \cup A_5))$, and $N_Q$ into Theorem 1, three quantities of interest are obtained by summing up indicators of subchains:

(i) $\pi_{N-1_A}$:

$$\frac{1}{\pi_{N-1_A}} = \frac{1}{\beta_{N-1_A,n-1_A} \cdot \pi_{n-1_A}} + \frac{1}{\pi_{N-1_A}} + \frac{1}{\beta_{N-1_A,n+1_B} \cdot \pi_{n+1_B}} + \frac{1}{\beta_{N-1_A,n+1_B} \cdot \pi_{n+1_B}}.$$

(ii) $P_B$:

$$\frac{P_B}{\pi_{N-1_A}} = \frac{1}{\pi_{N-1_A}} + \frac{1}{\beta_{N-1_A,n+1_B} \cdot \pi_{n+1_B}} + \frac{1}{\beta_{N-1_A,n+1_B} \cdot \pi_{n+1_B}}.$$
(iii) $L_Q$:  
\[
\frac{L_Q}{\pi_{N-1,A}^{0}} = E[N_Q] = \frac{L_Q^1}{\pi_{N-1,A}^{0}} + \frac{L_Q^2}{\pi_{N-1,A}^{0}} + \frac{L_Q^4}{\pi_{N-1,A}^{0}} + \frac{L_Q^5}{\pi_{N-1,A}^{0}}.
\]

Our next task is to analyze the five decomposed subchains. Derivation of closed-form solutions for the subchains is straightforward since these subchains can be analyzed independently by truncation (subchains $A_1$ and $A_3$) or termination (subchains $A_2$, $A_3$, and $A_4$) (see Appendix Section EC.3 for the derivation in detail). For notational convenience, let $X$ be a Poisson random variable with parameter $\lambda/\mu$: $E[X] = \text{var}(X) = \lambda/\mu$; $\Pr\{X = s\}$ and $\Pr\{X \leq s\}$ represent Poisson probability mass function and cumulative distribution function, respectively. We denote $s = c - e$, $\rho = \lambda/(s\mu)$, $\omega = 1/\rho = s\mu/\lambda$, and $\eta = \lambda/(\epsilon\mu)$. We assume that $n, N, c, e,$ and $s$ are all integers that satisfy $N > n > c$, $e > 0$, and $s = c - e > 0$. Define four functions as follows (see Appendix Section EC.3):

\[
f^1(k, \omega) = \frac{\Pr\{X \leq s\} \omega^{k-s}}{e_s} + \frac{1 - \omega^{k-s}}{1 - \omega}, \quad g^1(k, \omega) = \frac{(k-s) - (k-s+1)\omega + \omega^{k-s+1}}{(1-\omega)^2},
\]

\[
f^2(k, \omega) = \frac{k+1 - \omega(1 - \omega^{k+1})}{(1-\omega)^2}, \quad \text{and} \quad g^2(k, \omega) = \frac{k(k+1)}{2(1-\omega)} - \frac{(k - (k+1)\omega + \omega^{k+1})\omega}{(1-\omega)^3}.
\]

Performance indicators of subchains are summarized as follows:

(i) subchain $A_1$:

\[
\frac{1}{\pi_{N-1,A}^{0}} = f^1(n-1, \omega) \quad \text{and} \quad \frac{L_Q^1}{\pi_{N-1,A}^{0}} = g^1(n-1, \omega).
\]

(ii) subchain $A_2$:

\[
\frac{1}{\pi_{N-1,A}^{0}} = f^2(N-n-1, \omega) \quad \text{and} \quad \frac{L_Q^2}{\pi_{N-1,A}^{0}} = (n-s)f^2(N-n-1, \omega) + g^2(N-n-1, \omega).
\]

(iii) subchain $A_4$:

\[
\frac{1}{\pi_{N+1,B}^{0}} = f^2(N-n-1, \eta) \quad \text{and} \quad \frac{L_Q^4}{\pi_{N+1,B}^{0}} = (N-c)f^2(N-n-1, \eta) - g^2(N-n-1, \eta).
\]

(iv) subchain $A_5$:

\[
\frac{1}{\pi_{N+1,B}^{0}} = \frac{1}{1-\eta} \quad \text{and} \quad \frac{L_Q^5}{\pi_{N+1,B}^{0}} = \frac{N-c+1}{1-\eta} + \frac{\eta}{(1-\eta)^2}.
\]

We can also derive coefficients:

\[
\beta_{N-1,A,n-1,A} = \frac{1-\omega}{\omega(1-\omega^{N-n})}, \quad \beta_{N-1,A,n+1,B} = \frac{1}{\eta}, \quad \text{and} \quad \beta_{N-1,A,N+1,B} = \frac{1-\eta}{\eta^2(1-\eta^{N-n})}.
\]

Combining the above, we obtain the analytical representation for all necessary indicators to calculate the total cost $C_{staff} + C_{wait} + C_{switch}$:
(i) \( \pi_{N-1,A} \):
\[
\frac{1}{\pi_{N-1,A}} = \left( \frac{1}{1-\omega} + \frac{\eta}{1-\eta} \right)(N-n) - \frac{\omega^{n-s}(1-\omega^{N-n})}{1-\omega} \left( \frac{1}{1-\omega} - \frac{\Pr\{X \leq s\}}{\Pr\{X = s\}} \right).
\]

(ii) \( P_B \):
\[
\frac{P_B}{\pi_{N-1,A}} = \frac{\eta}{1-\eta} (N-n).
\]

(iii) \( L_Q \):
\[
\frac{L_Q}{\pi_{N-1,A}} = \left( \frac{1}{1-\omega} + \frac{\eta}{1-\eta} \right)(N-n) \left( \frac{N+n+1}{2} - \frac{\lambda}{\mu} - \frac{1}{1-\omega} + \frac{\eta}{1-\eta} \right).
\]

Figure 2  Total cost of operation as a function of \((n, N)\) thresholds with \(\lambda = 600/\text{hour}, \mu = 65/\text{hour}, c = 10, e = s = 5, c_g = 20/\text{hour}, c_w = 10/\text{hour}, \) and \(c_s = 50/\text{cycle}.\)

Using these expressions, we obtain a closed-form representation of the total cost, which is easy to evaluate in an Excel spreadsheet. Figure 2 is the exact cost surface for all possible \((n, N)\) combinations, from which we obtain the optimal CBS policy with thresholds \((n, N) = (10, 31)\) for the parameters given in the caption of Figure 2. This example illustrates the simplicity of our approach: We can derive analytical representations by a simple summation of performance indicators of subchains. For example, if we want to include customer abandonment in subchain 5 in this CBS model, we only need to replace the terms originated from subchain 5 with the new terms in the summation. Or if we consider adding a third threshold (or more), we only need to add terms corresponding to the subchains representing new thresholds in the summation: The complexity of the computation following our method is increased only linearly with respect to the number of subchains added.
5. Conclusions

We develop a new Markov chain (MC) decomposition method based on the total expectation theorem (the law of total expectation), one of the most utilized and powerful probability theorems. Because of the generality of the total expectation theorem, our method is very general: It allows us to decompose a MC into any collectively exhaustive subchains, and evaluate an expectation of any function of a Markov chain state (e.g., steady-state probabilities, first moment, second moment, etc.) Furthermore, a quantity of interest for the full MC is obtained as a simple sum of the quantities of subsets because the quantity of interest represented in our method satisfies all properties that a measure satisfies. Hence, for example, if subchains are overlapping, all we need to do is to subtract the contribution of the overlapping sets from the total sum. Using our method, we reveal previously unknown relationships among performance indicators of queueing systems (in Section 4.1); these are simply the identities of the total expectation theorem in MC settings. We also apply our method to derive performance indicators of a queue with adjustable staffing levels (in Section 4.2) – a complex queueing system – by a simple summation of properties of decomposed subchains. Each of these subchains could be replaced by other subchains if we want to extend the model to other similar models.

One key technique used in the paper is termination: The transitions among subchains are modified and redirected back to each subchain. This termination enables decomposed subchains to conserve partial flow at each boundary state, maintaining the same steady-state distribution as the original full MC, up to a normalization constant. We explain in the paper that there exists – and we can find – a termination scheme that works for any decomposed subchain.

Our method provides a potentially useful tool to find new properties in queueing systems; we can do so by decomposing the full MC representing the queueing system into subchains that are overlapping/nested and observe what identities hold as a consequence of the total expectation theorem. We can also apply our method to solve large, complex, MCs numerically. Two key benefits of our method in numerical applications are: (1) the ability to cope with MCs that grow or shrink, and (2) the ease of implementation in a parallel computation. For example, social networks observed in the Internet typically change their structures dynamically. Our algorithm can cope with such MCs that lose and add states at the same time: All we will need to do is to subtract and add performance indicators corresponding to those lost and added states, respectively. In addition, by allocating the calculation of each decomposed subchain to different CPUs, we can increase the speed of the calculation. We will explore the potential of our method in numerical applications in subsequent papers.
References

Bhandari, Atul, Alan Scheller-Wolf, Mor Harchol-Balter. 2008. An exact and efficient algorithm for the constrained dynamic operator staffing problem for call centers. *Management Science* **54**(2) 339–353.

Feinberg, Brion N, Samuel S Chiu. 1987. A Method to Calculate Steady-State Distributions of Large Markov Chains by Aggregating States. *Operations Research* **35**(2) 282–290.

Grimmett, Geoffrey, David Stirzaker. 2001. *Probability and Random Processes*. Oxford University Press.

Halmos, Paul Richard. 1950. *Measure Theory*. Van Nostrand.

Harchol-Balter, Mor. 2013. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press.

Katehakis, Michael N, Laurens C Smit. 2012. A Successive Lumping Procedure for a Class of Markov Chains. *Probability in the Engineering and Informational Sciences* **26**(04) 483–508.

Kelly, Frank P. 1979. *Reversibility and Stochastic Networks*. John Wiley & Sons.

Kemeny, John G, James Laurie Snell. 1960. *Finite Markov Chains*, vol. 356. van Nostrand Princeton, NJ.

Kleinrock, Leonard. 1975. *Queueing Systems, Volume I: Theory*. John Wiley & Sons.

Nelson, Randolph. 1995. *Probability, Stochastic Processes, and Queueing Theory*. Springer.

Simon, Herbert A, Albert Ando. 1961. Aggregation of variables in dynamic systems. *Econometrica: Journal of The Econometric Society* 111–138.

Stewart, William J. 1994. *Introduction to the Numerical Solution of Markov Chains*, vol. 41. Princeton University Press Princeton.

Takahashi, Yukio. 1975. A Lumping Method for Numerical Calculations of Stationary Distributions of Markov Chains. *Technical Report B-18, Department of Information Sciences, Tokyo Institute of Technology, Tokyo, Japan* .

Whittle, P. 1968. Equilibrium distributions for an open migration process. *Journal of Applied Probability* **5**(3) 567–571.

Whittle, Peter. 1986. *Systems in Stochastic Equilibrium*. John Wiley & Sons.

Zhang, Zhe George. 2009. Performance Analysis of a Queue with Congestion-Based Staffing Policy. *Management Science* **55**(2) 240–251.
Electronic Companion: Markov Chain Decomposition Based On Total Expectation Theorem

In this appendix, we derive analytical expressions for some of the performance indicators, which are used when we solve the MC representing the CBS model in Section 4.2. Note that these formulae can be obtained using the standard procedure for deriving steady-state probabilities and expectations. However, since an expectation $E[\cdot]$ itself is not what we need to know, but rather we seek a “scaled” expectation $E[\cdot]/\pi_k$, we follow our scheme to calculate quantities of interest in our “scaled” format. The benefit of this format is that we obtain a simpler analytical expression due to the elimination of a normalization constant. Following the paper, we use $N$ (or $N_Q$) to represent an operator for the number of vehicles/customers in the system (or queue) and $L$ (or $L_Q$) to represent its average in this appendix.

**EC.1. Performance Indicators for an M/M/s/k Queue**

We derive properties of interest (specifically, $1/\pi_s$ and $L_Q/\pi_s$) when subchains share a single state. Consider an M/M/s/k queue where $s$ is the number of servers and $k$ is the capacity of the system. Denote the arrival rate as $\lambda$ and the service rate for each server as $\mu$. We decompose the full MC $A = \{0, 1, 2, \cdots, k\}$ into two subchains sharing a single state $s$: $A_1 = \{0, 1, \cdots, s\}$ and $A_2 = \{s, s+1, \cdots, k-1, k\}$. Since subchains are connected at a single state, truncation is sufficient to conserve their steady-state distributions. Both truncated subchains are well-known queueing systems: $A_1$ is an M/M/s/s queue and $A_2$ is an M/M/1/k-s queue with a utilization parameter $\rho = \lambda/(s\mu)$.

To simplify the representation, let $X$ be a Poisson random variable with parameter $\lambda/\mu$: $E[X] = \text{var}(X) = \lambda/\mu$. Then, we know (for example, see Harchol-Balter 2013 and Kleinrock 1975)

$$\frac{1}{\pi_s} = \sum_{i=0}^{s} \frac{\lambda^i}{\mu^i i!} = \frac{\Pr\{X \leq s\}}{\Pr\{X = s\}}$$

and

$$\frac{1}{\pi_s^2} = 1 + \rho + \cdots + \rho^{k-s} = \frac{1 - \rho^{k-s+1}}{1 - \rho}.$$

Hence, using Equation (9), we obtain $1/\pi_k$ and $L_Q/\pi_k$:

$$\frac{1}{\pi_s} = \frac{1}{\pi_s^1} + \frac{1}{\pi_s^2} - 1 = \frac{\Pr\{X \leq s\}}{\Pr\{X = s\}} + \frac{\rho \cdot (1 - \rho^{k-s})}{1 - \rho}.$$

$$\frac{L_Q}{\pi_s} = \frac{L_Q^2}{\pi_s^2} = \frac{1 \cdot \pi_s^2 + 2 \cdot \pi_s^2 + \cdots + (k-s) \cdot \pi_s^2}{\pi_s^2} = \frac{1 \cdot \pi_s^2 + 2 \pi_s^2 + \cdots + (k-s) \cdot \pi_s^2}{\pi_s^2} = \rho \cdot (1 + 2 \rho + \cdots + (k-s) \rho^{k-s-1})$$
According to Equation (15), all we need to know is the expression for \( \frac{1}{\pi_k} \). Using this result and parameter \( \omega = 1/\rho \), we can convert the representation as follows:

\[
\frac{L_Q}{\beta_{k,s}} = \frac{L_Q}{\pi_k} = \frac{\rho \cdot [1 - (k - s + 1)\rho^{k-s} + (k - s)\rho^{k-s+1}]}{(1 - \rho)^2 \rho^{k-s}} = \frac{(k-s) - (k-s+1)\omega + \omega^{k-s+1}}{(1-\omega)^2} \equiv g^1(k, \omega).
\]

**EC.2. Performance Indicators for an M/M/1 Queue with Restart**

We consider a MC \( A_k = \{0, 1, 2, \cdots, k\} \), which is a birth and death MC with an extra transition (restart) from the last state \( k \) back to the first state \( 0 \). We are specifically interested in quantities \( 1/\pi_k^k \) and \( L^k/\pi_k^k \). A closed-form representation of these quantities is utilized in Section 4.2. According to Equation (15), all we need to know is the expression for \( 1/\beta_{k,i} = \pi_k^k/\pi_k^k = \beta_{i,k} \). Let \( \lambda_i \) be the arrival rate at state \( i \) (\( i = 0, 1, \cdots, k-1 \)) and \( \mu_i \) be the departure rate from state \( i \) (\( i = 1, 2, \cdots, k \)). Let \( r \) be the rate of transition from state \( k \) to state 0. Note that the flow balance equation holds: \( \lambda_i \pi_i^k = \mu_{i+1} \pi_{i+1}^k + r \pi_i^k \) for \( i = 0, 1, \cdots, k-1 \). Denote \( a_i = \mu_{i+1}/\lambda_i \) and \( b_i = r/\lambda_i \). Dividing the flow balance equation by \( \lambda_i \pi_i^k \), we obtain a recursive equation: \( \beta_{i,k} = a_i \beta_{i+1,k} + b_i \), where \( i = 0, 1, \cdots, k-1 \). By applying this recursive equation repeatedly with the condition \( \beta_{k,k} = 1 \), we can obtain the expression for \( \beta_{i,k} \).

In particular, in Section 4.2, we consider the simplest case where, for all \( i \), \( a_i = \omega \) (constant) and \( b_i = 1 \) hold. To derive \( 1/\pi_k^k \) and \( L^k/\pi_k^k \), we first find \( \beta_{i,k} \), which is simply expressed as

\[
\beta_{i,k} = 1 + \omega + \omega^2 + \cdots + \omega^{k-i} \beta_{i+(k-i),k} = \frac{1 - \omega^{k+1}}{1 - \omega}, \quad \text{or equivalently,} \quad \beta_{k-i,k} = \frac{1 - \omega^{i+1}}{1 - \omega}.
\]

In addition, we use the following property:

\[
\sum_{i=0}^{k} \omega^{i-1} = \sum_{i=0}^{k} \frac{\partial}{\partial \omega} \left( \sum_{i=0}^{k} \omega^i \right) = \frac{\partial}{\partial \omega} \left( \frac{1 - \omega^{k+1}}{1 - \omega} \right) = -\frac{(k+1)\omega^k}{(1-\omega)^2} + \frac{1 - \omega^{k+1}}{(1-\omega)^2}.
\]

By letting \( f(X) = 1 \) and \( N \) in Equation (15), we can obtain analytical expressions for \( 1/\pi_k^k \) and \( L^k/\pi_k^k \), respectively:

\[
\frac{1}{\pi_k} = \sum_{i=0}^{k} \beta_{i,k} = \sum_{i=0}^{k} \frac{1 - \omega^{i+1}}{1 - \omega} = \frac{1}{1 - \omega} \sum_{i=0}^{k} 1 - \frac{\omega}{1 - \omega} \sum_{i=0}^{k} \omega^i = \frac{k + 1}{1 - \omega} - \frac{\omega(1 - \omega^{k+1})}{(1-\omega)^2} = f^2(k, \omega).
\]
and

\[
\frac{L_k}{\pi_k} = \sum_{i=0}^{k} i \beta_{i,k} = \sum_{i=0}^{k} (k-i) \beta_{k-i,k} = \sum_{i=0}^{k} \frac{(k-i)(1-\omega^{i+1})}{1-\omega}
\]

\[
= \frac{k(k+1)}{2(1-\omega)} - \frac{k(\omega^{k+1})}{(1-\omega)^2} + \frac{(1-\omega^{k+1})}{(1-\omega)^3} \omega^2 = g^2(k,\omega).
\]

**EC.3. Performance Indicators of Subchains in CBS Model**

We decompose the full MC representing the CBS model into five (partially overlapping) subchains:

\(A_1 = \{0_A, 1_A, \cdots, n - 1_A\}, A_2 = \{n_A, n + 1_A, \cdots, N - 1_A\}, A_3 = \{N - 1_A, n + 1_B\}, A_4 = \{n + 1_B, n + 2_B, \cdots, N_B\}, A_5 = \{N + 1_B, N + 2_B, \cdots\}\). We denote \(s = c - e, \rho = \lambda/s\mu, \omega = 1/\rho = s\mu/\lambda, \) and \(\eta = \lambda/c\mu.\) We assume that \(n, N, c, e, \) and \(s\) are all integers that satisfy \(N > n \geq c, e > 0,\) and \(s = c - e > 0.\) We analyze each subchain independently.

1. **subchain \(A_1:** Since this subchain is connected to the rest at a single state, truncation is sufficient to conserve its steady-state distribution. A truncated subchain \(A_1\) is a regular \(M/M/s/k\) queue, whose solution is shown in Section [EC.1](#). Let \(X\) be a Poisson random variable with parameter \(\lambda/\mu: E[X] = \text{var}(X) = \lambda/\mu.\) By setting \(k = n - 1, s = c - e, \) and \(\omega = 1/\rho = s\mu/\lambda\) for the formulae for \(f^1(k,\omega)\) and \(g^1(k,\omega)\) in Section [EC.1](#), we obtain

\[
\frac{1}{\pi_{n-1_A}} = f^1(n-1,\omega) \quad \text{and} \quad \frac{L_k^1}{\pi_{n-1_A}} = g^1(n-1,\omega).
\]

2. **subchain \(A_2:** Since there is a single inflow state at state \(n_A,\) we can use Corollary [2](#) to determine the appropriate termination. A terminated subchain \(A_2\) is an \(M/M/1/k\) queue with restart. Notice that \(A_2\) starts from state \(n_A,\) where \(n - s(> 0)\) people are already in a queue. Hence, the average number of waiting people in \(A_2\) can be obtained by shifting the average number of people in the \(M/M/1/k\) queue by \(n - s.\) By setting \(k = (N - 1) - n = N - n - 1\) and \(\omega = 1/\rho = s\mu/\lambda\) for the formulae for \(f^2(k,\omega)\) and \(g^2(k,\omega)\) in Section [EC.2](#), we obtain

\[
\frac{1}{\pi_{N-1_A}} = f^2(N-n-1,\omega)
\]

and

\[
\frac{L_k^2}{\pi_{N-1_A}} = \frac{n-s}{\pi_{N-1_A}} + g^2(N-n-1,\omega) = (n-s)f^2(N-n-1,\omega) + g^2(N-n-1,\omega).
\]
Also, using the expression for $\beta_{i,k}$ in Section EC.2 by setting $i = 0$ and $k = N - n - 1$, we obtain
\[
\beta_{nA,N-1A} = \frac{1 - \omega^{N-n}}{1 - \omega}.
\]

(3) subchain $A_3$: Since there is a single inflow state $n + 1_B$ from chain $B$ and a single inflow state $N - 1_A$ from chain $A$, we can again use Corollary 2 to determine the appropriate termination. A terminated subchain $A_3$ is a two state MC, with a transition from $N - 1_A$ to $n + 1_B$ at rate $\lambda$ and a transition from $n + 1_B$ to $N - 1_A$ at a rate $c\mu$. We only need to know the $\beta$ coefficient for this subchain. Using $\eta = \lambda/(c\mu)$,
\[
\beta_{N-1A,n+1B} = \frac{1}{\eta}.
\]

(4) subchain $A_4$: This subchain is symmetric to subchain $A_2$. A terminated subchain $A_4$ is a reversed queueing system of the M/M/1/$k$ queue with restart in Section EC.2. Hence, the average number of people in $A_4$ can be obtained by subtracting the number in the original M/M/1/$k$ queue with restart from its capacity $k = N - n - 1$. As in (2) above, we need to shift the number by $n + 1 - c(\geq 0)$, who are already in a queue at the left-most state $n + 1_B$ in subchain $A_4$. Therefore, we obtain
\[
\frac{1}{\pi_{n+1B}^4} = f^2(N - n - 1, \eta)
\]
and
\[
\frac{L_Q^4}{\pi_{n+1B}^4} = \frac{k + (n + 1 - c)}{\pi_{N-1A}^2} - g^2(N - n - 1, \eta) = (N - c)f^2(N - n - 1, \eta) - g^2(N - n - 1, \eta).
\]

Also, using the expression for $\beta_{i,k}$ in Section EC.2 with the capacity $k = N - n - 1$ and the rate $\eta$, we obtain the following expression. (Note that the order of subscript in $\beta$ is reversed because we reverse the numbering of states in the MC in Section EC.2)
\[
\beta_{NB,n+1B} = \frac{1 - \eta^{N-n}}{1 - \eta}.
\]

(5) subchain $A_5$: Since this subchain is connected to the rest at a single state, truncation is sufficient to conserve its steady-state distribution. A truncated subchain $A_5$ is a regular M/M/1 queue with the utilization rate $\eta = \lambda/(c\mu)$, where the solution is well-known. The average waiting people is obtained by shifting the average number by $(N + 1) - c$, which is the number of waiting people at the left-most state $N + 1_B$ in subchain $A_5$. We obtain:
\[
\pi_{N+1B}^5 = 1 - \eta \quad \text{and} \quad L_Q^5 = N - c + 1 + \frac{\eta}{1 - \eta}.
\]

or equivalently,
\[
\frac{1}{\pi_{N+1B}^5} = \frac{1}{1 - \eta} \quad \text{and} \quad \frac{L_Q^5}{\pi_{N+1B}^5} = N - c + 1 + \frac{\eta}{(1 - \eta)^2}.
\]
The final task is to identify all $\beta$ coefficients. Notice that $\beta_{n_A,n-1_A} = 1/\omega$ and $\beta_{N_B,N+1_B} = 1/\eta$ hold. Hence, we can derive other necessary coefficients as follows:

$$\beta_{N-1_A,n-1_A} = \beta_{N-1_A,n_A} \cdot \beta_{n_A,n-1_A} = \frac{1 - \omega}{\omega(1 - \omega^{N-n})} \quad \text{and}$$

$$\beta_{N-1_A,N+1_B} = \beta_{N-1_A,n+1_B} \cdot \beta_{n+1_B,N_B} \cdot \beta_{N_B,N+1_B} = \frac{1 - \eta}{\eta^2(1 - \eta^{N-n})}.$$

**References**

Harchol-Balter, Mor. 2013. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press.

Kleinrock, Leonard. 1975. *Queueing Systems, Volume I: Theory*. John Wiley & Sons.