A tractable analytical model for large-scale congested protein synthesis networks

Carolina Osorio [1] Michel Bierlaire [2]

June 10, 2011

Abstract

This paper presents an analytical model, based on finite capacity queueing network theory, to evaluate congestion in protein synthesis networks. These networks are modeled as a set of single server bufferless queues in a tandem topology. This model proposes a detailed state space formulation, which provides a fine description of congestion and contributes to a better understanding of how the protein synthesis rate is deteriorated. The model approximates the marginal stationary distributions of each queue. It consists of a system of linear and quadratic equations that can be decoupled. The numerical performance of this method is evaluated for networks with up to 100,000 queues, considering scenarios with various levels of congestion. It is a computationally efficient and scalable method that is suitable to evaluate congestion for large-scale networks. Additionally, this paper generalizes the concept of blocking: blocking events can be triggered by an arbitrary set of queues. This generalization allows for a variety of blocking phenomena to be modeled.

1 Introduction

To synthesize proteins, the information of an mRNA (messenger RiboNucleic Acid) is translated. An mRNA consists of a strand of codons. The information of an mRNA is encoded in these codons (i.e. each codon codes for an amino acid) and is translated to form proteins using ribosomes as catalysts.

Protein synthesis involves three main phases: initiation, elongation and termination. These are depicted in Figure 1. This figure presents an mRNA strand that consists of a series of N codons , i.e. a set of codons in a tandem topology. Each codon is depicted by a vertical line on the mRNA. There are four ribosomes on the mRNA. Each ribosome is L codons long.

During the initiation phase, the ribosome binds to the mRNA at the first codon (or start codon). Then the ribosome advances along the mRNA one codon at a time. At each codon, elongation takes place. During elongation the corresponding codon (i.e. the underlying amino acid) is added to the growing protein chain. Termination occurs when the ribosome encounters the last codon (or termination codon). Both the ribosome and the newly formed protein are released, i.e. they unbind from the mRNA, and the ribosome is once again available for other translations.

For a given mRNA, the bound ribosomes advance along its codons, and may therefore be blocked by downstream ribosomes. Since for a given cell there are numerous mRNA's competing for available (i.e. non-binding) ribosomes, the blocking of ribosomes on an mRNA strand decreases the protein

[1] Civil and Environmental Engineering Department, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, 02139, USA, osorioc@mit.edu

[2] Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland, michel.bierlaire@epfl.ch

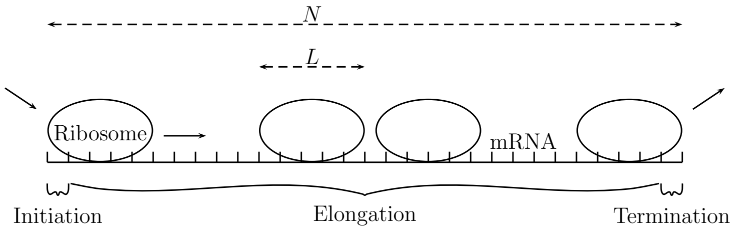

Figure 1: Ribosomes on an mRNA strand. Adapted from Mehra and Hatzimanikatis (2006).

synthesis rate of that mRNA, and may affect that of other mRNA's by reducing the probability that a ribosome is available for translation.

The frequency and effect of ribosome blocking is determined by the codon-specific initiation, elongation and termination rates, which therefore play an important role in the protein synthesis rate. Protein synthesis models are developed in order to study how these translation rates induce ribosome congestion and affect protein synthesis.

The main objectives and contributions of this paper are two-fold. First, to go beyond existing models by providing a more detailed description of ribosome congestion, which contributes to a better understanding of how the protein synthesis rate is deteriorated. Probabilistic protein synthesis models derive stationary distributions of the location of ribosomes along mRNA strands (see Mehra and Hatzimanikatis (2006) and references therein). In order to provide a more detailed description of ribosome (and codon) states, we use the *blocking* notion of finite capacity queueing theory, along with a detailed state space formulation to distinguish between active ribosomes and blocked ribosomes. This yields a more detailed quantification of congestion and its effects.

The second contribution is to enable the analysis of large-scale congested protein synthesis networks. An mRNA strand consists of a sequence of codons. In a small-genome organism the number of codons is of the order of 400,000 (Mehra and Hatzimanikatis, 2006). The study of protein synthesis involves large-scale networks. It requires scalable models that remain computationally efficient under congested conditions. The proposed model consists of a system of linear and quadratic equations. It is therefore particularly tractable and appropriate to address large-scale problems.

In this paper, we consider the ribosome congestion problem from a novel perspective, that of finite capacity queueing theory. This formulation is derived from a project in collaboration with the Laboratory of Computational Systems Biotechnology (LCSB) at Ecole Polytechnique Fédérale de Lausanne. Probabilistic analytical modeling of other intra-cell processes have been investigated by authors such as Gelenbe (2007, 2008).

In order to derive stationary distributions of the location of the different ribosomes along mRNA's, we proceed as in Mehra and Hatzimanikatis (2006) (hereafter referred to as the MH model). We describe the location of the ribosomes based on the location of their heads. Recall that a ribosome occupies L consecutive codons (Figure 1). The head of a ribosome refers to the part of the ribosome that occupies the most downstream of these L codons. Similarly to the MH model, we consider each codon and derive the stationary probability that there is a head of a ribosome at a given codon of an mRNA. Each codon is modeled as a single server bufferless queue. An mRNA consists of a series of codons, and is therefore modeled as a tandem network of single server bufferless queues.

This paper presents a general formulation that evaluates the impact of congestion for tandem single server bufferless networks. Since such networks are relevant for a variety of application fields, including manufacturing systems (Papadopoulos and Heavey, 1996; Dallery and Gershwin, 1992), computer systems (Balsamo et al., 2003) and telecommunication systems (Alfa and Liu, 2004; Artalejo, 1999), the numerical efficiency and scalability of this model is of wide interest.

Additionally, this paper defines a more general blocking concept. Traditionally, blocking at a given queue is triggered due to the state of the queues directly downstream. We allow blocking to be caused by the state of an arbitrary set of queues. This generalization allows for a variety of blocking phenomena to be modeled.

This paper is structured as follows. We review the analytical approaches for tandem networks of finite capacity queues (Section 2). The queueing model for bufferless single server tandem networks is presented in Section 3. The protein synthesis model is detailed (Section 4). We then use the model to evaluate ribosome congestion, and illustrate its performance for large-scale networks (Section 5).

2 Finite capacity queueing networks

In networks with finite capacity queues (i.e. queues with finite buffer sizes) the spread of congestion is modeled by what is known as blocking. A job is the generic name for the units of interest that flow through the network. In the application considered in this paper, the jobs are ribosomes. Blocking occurs when a job cannot proceed to the next queue on its path because that queue is full. The job is said to be blocked at its current location. Various types of blocking mechanisms have been defined (Balsamo et al., 2001).

These blocking mechanisms lead to complex between-queue dependencies. Describing this blocking phenomenon (i.e. where and how often it occurs, as well as its duration) analytically is challenging; not to mention the added complexity of deriving a computationally efficient model. The analytical analysis of finite capacity queueing networks (FCQN) is intricate, and is therefore limited to the stationary regime.

An introductory book to FCQN is Balsamo et al. (2001). Several reviews and historical overviews of FCQN methods exist (Perros, 2003; Balsamo et al., 2003; Artalejo, 1999; Papadopoulos and Heavey, 1996; Perros, 1984). Exact methods to evaluate the stationary performance measures of FCQN exist only for tandem networks with two queues (e.g. Grassman and Derkic, 2000; Akyildiz and von Brand, 1994). In order to evaluate the performance of larger networks, approximation methods are developed.

Approximation methods may be either analytical or simulation-based. Here we consider analytical models. Dallery and Frein (1993) present a review of analytical approximate methods for tandem finite capacity networks with exponentially distributed service times. They also propose a classification of these methods.

In order to reduce the dimensionality and complexity of analytically analyzing FCQN, approximation methods decompose the network into subnetworks. Each subnetwork is then analyzed independently, yielding performance measures at the subnetwork level. Existing approaches for tandem networks have decomposed the network into subnetworks of one or two queues.

Decomposing the network into single queues is the most common approach to analyze FCQN. Methods for tandem networks include Jun and Perros (1990), Altiok (1989) and Altiok (1982). These three methods present numerical results for networks with up to six queues. A method developed to address larger tandem networks is presented in Gershwin (1987). The numerical examples include instances with up to 20 queues. The Expansion Method (Kerbache and Smith, 2000) has been used for tandem networks (Cruz et al., 2005). Considering more general topologies, it has been used to address larger networks allowing for 70 queues (Kerbache and Smith, 2000). Single queue decomposition methods have been investigated for feed-forward topologies with up to 630 queues (Osorio and Bierlaire, 2009), and for tandem topologies with 144 queues (Osorio, 2010).

Two-queue decomposition methods derive stationary performance measures for pairs of queues. Various two-queue decomposition methods for open tandem networks have been proposed (e.g. van Vuuren et al., 2005; Alfa and Liu, 2004). Such methods yield marginal distributions for pairs of queues, rather than single queues, and can therefore lead to more accurate results. Nevertheless, they are computationally more demanding (Perros, 1994).

Most analytical approximation methods for tandem networks have limited their analysis to networks with less than 100 queues. We are interested in large-scale networks, with several thousand queues. This paper proposes a model for single server bufferless queues in a tandem topology. The model consists of a computationally tractable set of linear and quadratic equations. Such a formulation is scalable, and enables us to evaluate the performance of large-scale congested networks.

3 Model

The queueing model proposed in this work builds upon the model presented in Osorio and Bierlaire (2009), which is referred to as the *base model*. In this section, we introduce the assumptions and notations of the base model that are of interest for the current framework. We then prove that for single server tandem bufferless network, the system of equations of the base model is equivalent to a system of linear and quadratic equations that can be decoupled, leading to a tractable and scalable model. For a more detailed description and derivation of the base model, we refer the reader to Osorio and Bierlaire (2009).

3.1 Base model

3.1.1 Describing congestion through blocking

The base model considers a network of multiple-server queues in an arbitrary topology network. The main feature of the base model is the explicit modeling of the blocking phase. We use the blocking mechanism known as *blocking-after-service*, where blocking occurs as follows. A job:

1. arrives to a queue,
2. waits if all the servers are occupied,
3. is served (this is called the active phase),
4. is blocked if the next queue on its path is full (this is called the blocking phase),
5. leaves the queue.

The blocking phase is explicitly modeled via a novel formulation of the state space. The state of queue i is described by the number of active jobs A_i, blocked jobs B_i and waiting jobs W_i. Thus, the stationary marginal distribution of queue i is given by the probabilities: $P(A_i = a, B_i = b, W_i = w)$, for all feasible triplets (a, b, w).

Other finite capacity queueing models derive stationary marginal queue length distributions, i.e. they yield the probabilities $P(A_i+B_i+W_i = a+b+w)$. The base model derives marginal distributions that distinguish between active and blocked states. This allows for a detailed description of blocking and congestion.

3.1.2 Structural parameters

In order to approximate the stationary marginal distribution of the queues, the base model introduces a set of structural parameters that approximately capture the between-queue interactions. We first introduce their notation, we then detail their structural interpretation and present the corresponding equations. The index i refers to a given queue.

γ_i	external arrival rate;
λ_i	total arrival rate;
λ_i^{eff}	effective arrival rate;
μ_i	service rate;
$\tilde{\mu}_i$	unblocking rate;
μ_i^{eff}	effective service rate;
$\mathcal{P}_i$	probability of being blocked at queue i;
t_i	probability that queue i is not full.

Arrivals Three types of arrival rates are considered. Arrivals that arise from outside of the network are called external arrivals, they arise to queue i with rate γ_i. The total arrival rate, λ_i, accounts for both internal arrivals (that arise from upstream queues) and external arrivals. In the base model, all external arrivals that arrive to queue i while queue i is full are assumed to be lost. This type of queueing models are known as *loss models*. This leads to an *effective arrival rate* λ_i^{eff}, which accounts only for the arrivals that are actually processed, i.e. it excludes all lost arrivals.

Service, blocking and unblocking Blocking at a given queue i is described by two main parameters that approximate its occurrence and its duration. The first is captured by the probability with which a job at queue i is blocked $\mathcal{P}_i$. The second is captured by the unblocking rate $\tilde{\mu}_i$.

The probability that a queue is full corresponds to the probability that it will block upstream jobs. In finite capacity queueing theory, this probability is known as the *blocking probability*. Here it is given by $1-t_i$. Thus, $\mathcal{P}_i$ is determined by the blocking probabilities of the downstream queues of queue i.

A job is served (with rate μ_i), it is blocked (with probability $\mathcal{P}_i$) and is eventually unblocked (with rate $\tilde{\mu}_i$). The *effective service rate* of queue i, μ_i^{eff}, accounts for both service and blocking.

The base model approximates these structural parameters. For instance, the total and the effective arrival rates to a given queue are a function of upstream arrival rates. Similarly, the effective service rate takes into account blocking due to downstream queues. These structural parameters are used, along with the global balance equations, to approximate the marginal distributions.

To ensure tractability, the base model resorts to classical distributional assumptions and approximations. For each queue, the base model assumes independent and exponentially distributed service times. The times between successive arrivals and unblockings are approximated as independent and exponentially distributed random variables. More details concerning the distributional assumptions and approximations are given in Osorio (2010).

The base model was validated by comparison with exact results, simulation results and existing methods on networks with various topologies, including tandem topologies, and varying scenarios, namely under high intensity traffic (Osorio, 2010; Osorio and Bierlaire, 2009).

3.2 Arbitrary blocking structure

The traditional concept of blocking, assumes that blocking events at queue i are triggered by queues that are directly downstream, i.e. by any queue j such that a transition from queue i to queue j can take place. Recall that a ribosome is L codons long (Figure 1). If there is a head of a ribosome at a given codon i, then blocking can occur if there is a head of a ribosome L codons downstream. In order to capture this type of blocking, we generalize the blocking concept captured by the base model. We allow for blocking events to be triggered due to an arbitrary queue being full. We introduce the following notation:

$\mathcal{D}_i$ set of downstream queues of queue i;
$\mathcal{T}_i$ set of queues that can trigger blocking at queue i (refered to as trigger queues).

The set of downstream queues, $\mathcal{D}_i$, consists of the set of queues from which a transition from queue i can take place. It is determined by the transition probabilities:

$$\mathcal{D}_i = \{j, p_{ij} > 0\}. \tag{1}$$

For each queue j in $\mathcal{D}_i$, the probability that it is chosen is given by p_{ij}. Similarly, for a given queue j in $\mathcal{T}_i$, we denote q_{ij} the probability that it is chosen. The elements $\mathcal{D}_i, \mathcal{T}_i, (p_{ij})$ and (q_{ij}) allow us to generalize the blocking concept.

In the traditional blocking concept, a job at queue i can be blocked by its downstream queues, i.e. the set of trigger queues consists of the set of downstream queues:

$$\begin{cases} \mathcal{T}_i = \mathcal{D}_i & \text{(2a)} \\ q_{ij} = p_{ij}, \ \forall j \in \mathcal{T}_i. & \text{(2b)} \end{cases}$$

In the protein synthesis case, a job at queue i can be blocked if queue $i + L$ is full. This corresponds to:

$$\mathcal{T}_i = \{i + L\} \tag{3}$$

$$q_{ij} = \begin{cases} 1 & \text{if } j = i + L, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

3.3 Single server networks

We model each codon of an mRNA as a single server queue. The system of equations of the base model applied to single server queues and allowing for an arbitrary blocking structure is given by:

$$\begin{cases} \pi(i)g(\lambda_i, \mu_i, \tilde{\mu}_i, \mathcal{P}_i) = 0 & \text{(5a)} \\ \lambda_i = \lambda_i^{\text{eff}}/t_i & \text{(5b)} \\ \lambda_i^{\text{eff}} = \gamma_i t_i + \sum\limits_{j \in \mathcal{D}_i} p_{ji} \lambda_j^{\text{eff}} & \text{(5c)} \\ \dfrac{1}{\mu_i^{\text{eff}}} = \dfrac{1}{\mu_i} + \mathcal{P}_i/\tilde{\mu}_i & \text{(5d)} \\ \dfrac{1}{\tilde{\mu}_i} = \sum\limits_{j \in \mathcal{T}_i} \dfrac{\sum_k \widetilde{q_{kj}} \lambda_k^{\text{eff}}}{\lambda_i^{\text{eff}} \mu_j^{\text{eff}}} & \text{(5e)} \\ \mathcal{P}_i = \sum\limits_{j \in \mathcal{T}_i} q_{ij}(1 - t_j), & \text{(5f)} \\ \tilde{q}_{ij} = q_{ij}(1 - t_j)/\mathcal{P}_i, & \text{(5g)} \end{cases}$$

where $\pi(i)$ represents the stationary marginal distribution of queue i, and $(\tilde{q}_{ij})$ are the trigger probabilities conditional on a job being blocked at queue i (i.e. given that a job is blocked at queue i, $\tilde{q}_{ij}$ represents the probability that it is blocked by queue j).

We summarize the main features of these equations, and refer the reader to Osorio and Bierlaire (2009) for a detailed description. Equation (5a) corresponds to the global balance equations, it determines the marginal distribution for queue i. All other equations approximate the structural parameters of queue i.

The total and effective arrival rates are given by combining flow conservation with loss model information (Equations (5b) and (5c)) The effective service rate (Equation (5d)) is approximated as a function of the service rate, the probability that blocking occurs and the unblocking rate.

Blocking at queue i is described by the probability with which it occurs, as well as the rate with which it dissipates. The probability with which blocking occurs at queue i, $\mathcal{P}_i$, is determined by the blocking probabilities of its trigger queues (Equation (5f)). This equation states that a given trigger queue j is chosen among $\mathcal{T}_i$ with probability q_{ij}, once chosen it triggers blocking if it is full with probability $1 - t_j$. Equation (5e) approximates the unblocking rate $\tilde{\mu}_i$ at queue i. A detailed description of how this equation is derived is provided in the Appendix.

This system of equations is valid for single server queues with an arbitrary finite capacity, organized in an arbitrary topology network. For each queue, the exogenous parameters are (p_{ij}), (q_{ij}), μ_i and γ_i. All other variables are endogenous.

3.4 Tandem network model

Since each mRNA is modeled as a set of queues in a tandem topology, hereafter we consider a network of N queues in a tandem topology. The queues are indexed 1 to N, where queue 1 is the most upstream and queue N the most downstream. We assume that external arrivals only arise at the first queue, and that departures only occur at the last queue. This corresponds to transition probabilities given by:

$$p_{ij} = \begin{cases} 1 & \text{if } i < N \text{ and } j = i + 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

In tandem networks with a classical blocking structure, a job is blocked at queue i if upon service completion queue $i + 1$ is full. In this paper we consider that a job is blocked at queue i if upon service completion queue $i + L$ is full. The classical setting can therefore be retrieved by setting L equal to 1.

Queues that cannot be blocked are referred to as *terminal* queues, as opposed to *non-terminal* queues. Terminal queues correspond to queues such that the set $\mathcal{T}_i$ is empty. For the considered scenario where $\mathcal{T}_i = \{i + L\}$, the terminal queues are the queues indexed $N - L + 1$ to N, whereas all other queues are non-terminal.

In the next two sections, we show how the System of Equations (5) simplifies for tandem topology networks. We first present the equations for the structural parameters (Equations (5b)-(5g)), we then detail the global balance equations (Equation (5a)).

3.4.1 Structural parameters

External arrivals arise only at the first queue with rate γ, i.e.:

$$\gamma_i = \begin{cases} \gamma & \text{if } i = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Inserting Equations (6) and (7) into (5c) yields:

$$\begin{cases} \lambda_i^{\text{eff}} = \lambda_{i-1}^{\text{eff}}, \forall i > 1 & \text{(8a)} \\ \lambda_1^{\text{eff}} = \gamma t_1. & \text{(8b)} \end{cases}$$

That is, the effective arrival rate is constant across queues. Equation (5b) yields:

$$\forall i \ \lambda_i^{\text{eff}} = \lambda_i t_i = \text{constant}, \tag{9}$$

where the constant is given by Equation (8b):

$$\forall i \ \lambda_i^{\text{eff}} = \lambda_i t_i = \gamma t_1. \tag{10}$$

We first present the system of equations for the structural parameters, we then detail their derivation.

Terminal queues

$$\forall i \in [N - L + 1, N], \begin{cases} \lambda_i^{\text{eff}} = \gamma t_1 & \text{(11a)} \\ \mu_i^{\text{eff}} = \mu_i & \text{(11b)} \\ \dfrac{1}{\tilde{\mu}_i} = 0 & \text{(11c)} \\ \mathcal{P}_i = 0 & \text{(11d)} \end{cases}$$

Non-terminal queues

$$\forall i \in [1, N - L], \begin{cases} \lambda_i^{\text{eff}} = \gamma t_1 & \text{(12a)} \\ \dfrac{1}{\mu_i^{\text{eff}}} = \dfrac{1}{\mu_i} + (1 - t_{i+L}) \dfrac{1}{\mu_{i+L}^{\text{eff}}} & \text{(12b)} \\ \tilde{\mu}_i = \mu_{i+L}^{\text{eff}} & \text{(12c)} \\ \mathcal{P}_i = 1 - t_{i+L}. & \text{(12d)} \end{cases}$$

Let us show that the Systems (11)-(12) are equivalent to the system of equations for the structural parameters of the base model (Equations (5b)-(5g)). Equations (11a) and (12a) result from Equation (10). The System (11) concerns terminal queues. By definition these queues cannot be blocked, thus their expected blocked time (Equation (11c)) and their probability of being blocked (Equation (11d)) are null. Their conditional trigger probabilities are not defined (Equation (5g)). Inserting Equations (11c) and (11d) in (5d) yields Equation (11b).

For non-terminal queues, inserting Equations (3) and (4) into (5f) yields (12d). By inserting Equations (3), (4) and (5f) into (5g) leads to:

$$\tilde{q}_{ij} = \begin{cases} 1 & \text{if } j = i + L, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Inserting Equations (3), (10) and (13) into (5e) yields (12c). Equation (12b) is obtained by inserting Equations (12c) and (12d) into (5d).

3.4.2 Global balance equations

Recall that each codon is modeled as single server queue with no buffer. So far, the assumption of bufferless queues was not necessary. All previous equations are therefore valid for single server queues with an arbitrary finite capacity. To simplify the global balance equations, we will use the bufferless assumption. Hereafter, we assume bufferless queues.

Recall from Section 3.1.1 that the state space of each queue is defined as the sample space of the triplet of random variables (A_i, B_i, W_i), where A_i, B_i and W_i denote respectively the number of active, blocked and waiting jobs at queue i.

For single server bufferless queues the state space consists of three states. A queue can be in one of the following three states:

- empty: $(A_i, B_i, W_i) = (0, 0, 0)$,
- active: $(A_i, B_i, W_i) = (1, 0, 0)$, (i.e. the server of the queue is occupied by an active job),
- blocked: $(A_i, B_i, W_i) = (0, 1, 0)$, (i.e. the server of the queue is occupied by a blocked job).

We denote the probability of these three states as follows:

t_i	probability that queue i is empty;
y_i	probability that queue i is blocked;
z_i	probability that queue i is active.

The marginal distribution of queue i is given by: $\pi(i) = (t_i, y_i, z_i)$. Note that t_i is also the probability that the queue is not full, whereas the blocking probability (i.e. the probability that the queue is full) is given by $y_i + z_i$. We show that for single server bufferless queues in a tandem topology the global balance equations given by Equation (5a) lead to the following systems of equations:

Terminal queues

$$\forall i \in [N-L+1, N], \begin{cases} t_i + z_i = 1 & (14a) \\ y_i = 0 & (14b) \\ \mu_i z_i = \gamma t_1. & (14c) \end{cases}$$

Non-terminal queues

$$\forall i \in [1, N-L], \begin{cases} t_i + y_i + z_i = 1 & (15a) \\ y_i = (y_{i+L} + z_{i+L})^2 & (15b) \\ \mu_i z_i = \gamma t_1. & (15c) \end{cases}$$

Let us detail how these equations are derived. In the case of terminal queues the global balance equations are defined by:

$$\begin{cases} t_i + z_i = 1 & (16a) \\ y_i = 0 & (16b) \\ \lambda_i t_i - \mu_i z_i = 0. & (16c) \end{cases}$$

Note that Equation (16c) balances arrival and service events, and Equation (16b) states that terminal queues cannot be blocked. Since $\forall i \; \lambda_i t_i = \gamma t_1$ (Equation (10)), then the Systems of Equations (14) and (16) are equivalent. In the case of non-terminal queues, the global balance equations are defined by:

$$\begin{cases} t_i + y_i + z_i = 1 & \text{(17a)} \\ -\tilde{\mu}_i y_i + \mathcal{P}_i \mu_i z_i = 0 & \text{(17b)} \\ \lambda_i t_i - \mu_i z_i = 0. & \text{(17c)} \end{cases}$$

Note that Equation (17b) balances blocking and unblocking events, while Equation (17c) balances arrival and service events. Let us show that the Systems (15) and (17) are equivalent.

As for terminal queues, we use Equation (10) to obtain the equivalence between Equations (15c) and (17c). Thus, to show that the Systems (15) and (17) are equivalent we need to show the equivalence between Equations (15b) and (17b). To do so, we use the following lemma.

Lemma 1 *Let $H(i)$ denote the hypothesis that Equation* (15b) *holds for queue i. $H(N-L)$ holds, and if $H(k)$ holds $\forall k \in [i+1, N-L]$, then $H(i)$ holds.*

To prove this lemma we proceed by recursion. The proof is given in the Appendix. Thus, the global balance equations for single server bufferless queues in a tandem topology are given by the Systems (14) and (15).

The initial formulation of the global balance equations (Systems (16) and (17)) involves four structural parameters: $\lambda_i, \mu_i, \tilde{\mu}_i$ and $\mathcal{P}_i$. The parameter μ_i is exogenous, the other three parameters are endogenous. The equivalent formulation that we have just derived (Systems (14) and (15)) no longer involves any endogenous structural parameters. Thus, the system of equations can be decoupled. We can solve Systems (14) and (15) to obtain the marginal distributions for each queue, and if the structural parameters are of interest we can then solve Systems (11) and (12).

The system of equations for the structural parameters consists of $4N-2L$ equations, of which $3N-L$ are linear, and the remaining $N-L$ are quadratic. The global balance equations consist of $3N-L$ equations, of which $2N$ are linear and the remaining $N-L$ are quadratic.

4 Protein synthesis network

In this section, we build upon the model for tandem single server bufferless networks of Section 3.4 to derive the protein synthesis network model. The main aspects of protein synthesis that we are interested in modeling are described in Section 1. For a more detailed description of the protein synthesis process see Mehra and Hatzimanikatis (2006). We follow the same reasoning and assumptions as the MH protein synthesis model.

4.1 Stationary distributions

The MH model is an analytical codon-scale model of the translation of mRNAs into proteins. This model explicitly describes the phases of initiation, elongation and termination, and yields stationary distributions for each codon. A slightly modified version of this model is presented in Mier-y-Teran-Romero et al. (2009).

As described in Section 1, we proceed similarly to the MH model. We model each codon and yield the stationary probabilities that it is occupied by the head of a ribosome. Each codon is modeled as a bufferless queue with one server. Thus, an mRNA consists of a network of single server bufferless queues in tandem.

The main novelty of the MH method is to account for the blocking of ribosomes. The MH model does this by reducing the elongation rates using the conditional probability that codon $i+1$ is empty given that codon i is occupied. This conditional probability is used to approximate the blocking probability.

We provide a more detailed description of ribosome blocking, by considering that each codon can be in one of three states:

- the codon is occupied by the head of an *active* ribosome
- the codon is occupied by the head of a *blocked* ribosome
- the codon is not occupied by the head of a ribosome, i.e. it is either not covered by a ribosome at all, or it is covered by a part of the ribosome that is not the head.

In other words, given that a codon is occupied by the head of a ribosome, we distinguish between whether the ribosome is active or blocked.

The marginal stationary distribution is composed of:

t_i : the probability that codon i is not occupied by the head of a ribosome;
y_i : the probability that codon i is occupied by the head of a *blocked* ribosome;
z_i : the probability that codon i is occupied by the head of an *active* ribosome.

By modeling the ribosome blocking problem with a finite capacity queueing approach that distinguishes between active and blocked queues, we provide a finer description of ribosome blocking, and thus a more detailed quantification of congestion along an mRNA strand. This more disaggregate state space formulation can lead to a better understanding of how ribosome congestion affects the protein synthesis rate.

4.2 Structural parameters

In this protein synthesis context, there are two differences with the model presented in Section 3.4

1. There is a fixed and limited number of ribosomes that can bind to the mRNAs. The external arrival rate, γ, is therefore a function of the expected number of available (i.e. non-binding) ribosomes. It is no longer exogenous. This model therefore assumes a finite population of jobs (i.e. ribosomes). The approximation for γ is based on that of the MH model:

$$\gamma = a_0 + a_1 \sum_{i=1}^{N} (1 - t_i). \tag{18}$$

   Equation (18) concerns the external arrival rate, and approximates it as a function of the number of available (i.e. non-binding) ribosomes and of two exogenous parameters a_0 and a_1. The expression is taken from Equations (5) and (6) of Mehra and Hatzimanikatis (2006), or equivalently from their scaled versions which appear as Equation (11a) in Mier-y-Teran-Romero et al. (2009).

2. To start the translation process, a ribosome binds to the first codon of the mRNA. If this first codon is full, the ribosome cannot bind. Since a ribosome covers L consecutive codons, the

first codon is full if there is a head of a ribosome on any of the *first L codons*. These first L codons are called the *initiation site*. The probability that the first codon is free, i.e. that the initiation site does not contain a head of a ribosome, is denoted w_1. Its approximation is based on that of the MH model (Equation (5) in that paper), and is given by:

$$w_1 = 1 - \sum_{i=1}^{L} (1 - t_i). \tag{19}$$

The model of Section 3.4 assumes that external arrivals may be lost if the first queue is full. In this protein synthesis application, we assume that external arrivals may be lost if the initiation site is full. Thus Equations (11a) and (12a) become:

$$\lambda_i^{\text{eff}} = \gamma w_1. \tag{20}$$

That is, the effective arrival rate is now a function of the probability that the initiation site is free w_1, rather than the probability that the first queue is free t_1.

4.3 System of equations

The system of equations is given by:

$$\forall i \in [N-L+1, N], \begin{cases} t_i + z_i = 1 & \text{(21a)} \\ y_i = 0 & \text{(21b)} \\ \mu_i z_i = \gamma w_1 & \text{(21c)} \end{cases}$$

$$\forall i \in [1, N-L], \begin{cases} t_i + y_i + z_i = 1 & \text{(22a)} \\ y_i = (y_{i+L} + z_{i+L})^2 & \text{(22b)} \\ \mu_i z_i = \gamma w_1. & \text{(22c)} \end{cases}$$

$$\begin{cases} w_1 = 1 - \displaystyle\sum_{i=1}^{L} (1 - t_i) & \text{(23a)} \\ \gamma = a_0 + a_1 \displaystyle\sum_{i=1}^{N} (1 - t_i). & \text{(23b)} \end{cases}$$

$$\forall i \in [N-L+1, N], \begin{cases} \lambda_i^{\text{eff}} = \gamma w_1 & \text{(24a)} \\ \mu_i^{\text{eff}} = \mu_i & \text{(24b)} \\ \dfrac{1}{\tilde{\mu}_i} = 0 & \text{(24c)} \\ \mathcal{P}_i = 0 & \text{(24d)} \end{cases}$$

$$\forall i \in [1, N-L], \begin{cases} \lambda_i^{\text{eff}} = \gamma w_1 & \text{(25a)} \\ \dfrac{1}{\mu_i^{\text{eff}}} = \dfrac{1}{\mu_i} + (1 - t_{i+L}) \dfrac{1}{\mu_{i+L}^{\text{eff}}} & \text{(25b)} \\ \tilde{\mu}_i = \mu_{i+L}^{\text{eff}} & \text{(25c)} \\ \mathcal{P}_i = 1 - t_{i+L}, & \text{(25d)} \end{cases}$$

The Systems (21), (22), (24) and (25) are obtained by combining Equation (20) and the Systems (11)-(15). The System (23) is given by Equations (18) and (19).

Here, the only exogenous parameter is μ_i, all other variables are endogenous. These systems can be decoupled. In particular, if the parameters of interest are the stationary distributions, it is sufficient to implement Equations (21), (22) and (23). In this case, for a set of N codons, the system of equations consists of $3N + 2$ equations. There are $N + L + 2$ linear equations and $2N - L$ quadratic equations. Let us compare the main properties of this formulation to that of the MH model.

Congestion decomposition One of the contributions of the MH model is to acknowledge the interactions between the initiation, elongation, termination and protein synthesis rates. In particular, given a set of ribosomes on an mRNA, the model acknowledges that their translation rate may be deteriorated by the presence of downstream ribosomes, that prevent the ribosome from advancing (this is captured by the fraction of Equation (8) of the MH model).

We go beyond this by describing these ribosome congestion effects in more detail. By using the *blocking* phenomenon of finite capacity queueing theory, and the detailed state space formulation of the finite capacity queueing model, we disaggregate the state "a codon is occupied" into two states "occupied and blocked" and "occupied and active". By distinguishing between active and blocked codons, we provide more detailed distributional estimates. Additionally, the endogenous parameters of the proposed model provide a fine decomposition of congestion (e.g. in terms of its sources, frequency, impact).

Computational efficiency To evaluate the stationary distributions of each codon, two procedures have been used in Mehra and Hatzimanikatis (2006) and in Mier-y-Teran-Romero et al. (2009). The first solves a bilevel nonlinear optimization problem, the second solves a system of ordinary differential equations. The procedure proposed in this section consists of a system of linear and quadratic equations, its implementation is straightforward, and it can be solved with less complex numerical methods.

Nonetheless, if transient distributions are of interest these can only be derived by the method of Mier-y-Teran-Romero et al. (2009).

Scalability The number of equations that need to be implemented can be substantially reduced by identifying the queues that have equal service rates, μ_i. In this case, Equations (21c) and (22c) indicate that these queues also have a common value for z_i, since

$$z_i = \gamma w_1/\mu_i = \text{constant}. \tag{26}$$

If among the N queues there are D distinct service rates, then the number of equations reduces to $2N - L + D + 2$. In the case of protein synthesis, this can occur if the codons have common elongation rates. There are then three distinct service rates: initiation rate, termination rate and elongation rate, and the number of equations becomes $2N - L + 5$.

5 Empirical Analysis

5.1 Ribosome blocking

In this section, we use the protein synthesis model (Equations (21)-(25)) to evaluate ribosome congestion. The exogenous parameters of the queueing model are calibrated based on typical translation parameters provided by the members of the Laboratory of Computational Systems Biotechnology. We consider a single mRNA species with 144 codons, and assume that each ribosome covers 12 codons, i.e. $N = 144,\ L = 12$, as in Mier-y-Teran-Romero et al. (2009).

We solve the system of equations for all queues simultaneously, with the Matlab routine for non-linear systems of equations, *fsolve* (Mathworks, Inc., 2008). For a given tolerance, *tol*, convergence is attained when either the absolute values of all equations are smaller than *tol* or when both the sum of squares of the system of equations is smaller than $\sqrt{tol}$ and the change of its relative value is smaller than $max(tol^2, eps)$, where *eps* is the machine precision which is of magnitude 10^{-16}. The tolerance is chosen as $tol = 10^{-6}$. This choice is based on the criteria given in Dennis and Schnabel (1996). The distributions are initialized using uniform distributions.

We consider a set of scenarios with fixed initiation and elongation rates, and increasing termination rates. In queueing theory terms, this corresponds to the fixed service rates for all but the most downstream queue (indexed N), and increasing service rate for queue N.

Figure 2 displays for the different scenarios the probability that there is a head of a ribosome for the most downstream quarter of queues (indexed 109 to 144). For a given codon i, this probability is given by $y_i + z_i$.

The first scenario clearly illustrates how the model captures ribosome blocking. This scenario is the one with the smallest termination rate. Its probabilities are displayed with crosses. Since the termination rate is the limiting factor, the probability that a ribosome head remains at codon N (i.e. codon 144) is high. This leads to blocking L codons upstream, i.e. a high probability for codon 132. This blocking also propagates $2L$ codons upstream to codon 120.

As the termination rate increases, these probabilities decrease. For scenario 2, the impact of the termination rate on the occupation of codon 120 is low, yet the occupation of codons 132 and 144 remains high. For scenarios 3-5, the impact on codon 132 decreases. As the termination rate increases, the occupation probabilities of codons $120, 132$ and 144 (i.e. $N - 2L, N - L$ and N) decreases. For all scenarios, the computation time needed to evaluate the model is less than 0.3 seconds.

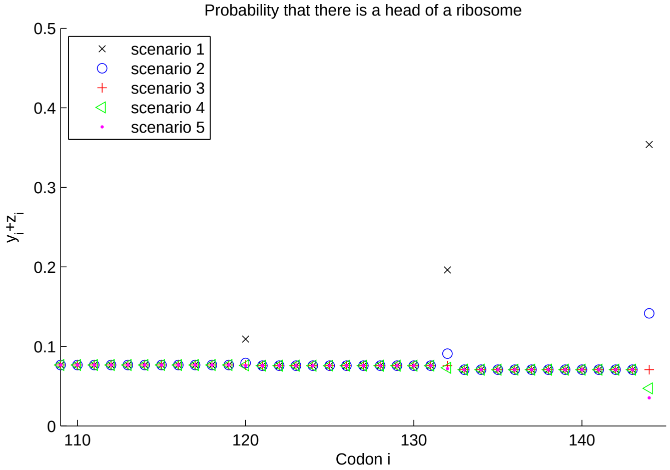

Figure 2: Probability that a codon is occupied by the head of a ribosome for scenarios with increasing termination rates.

We now show how the queueing model disaggregates these codon occupation probabilities, into "occupied and active" versus "occupied and blocked". Figure 3 considers the same scenarios and codons. The left plot of Figure 3 displays the probabilities that a codon is occupied by the head of an active ribosome. For a given codon i, this is denoted in the model by z_i. This plot shows that as the termination rate increases, only the probabilities of the last codon are affected, z_N.

The right plot presents the probabilities that a codon is occupied by the head of a blocked ribosome. For a given codon i, this is denoted in the model by y_i. This plot shows that the codons indexed 120 and 134 have large blocking probabilities. These decrease as the termination rate increases. Note that codons indexed $135 - 144$ have null probabilities, since they are terminal codons and cannot be blocked.

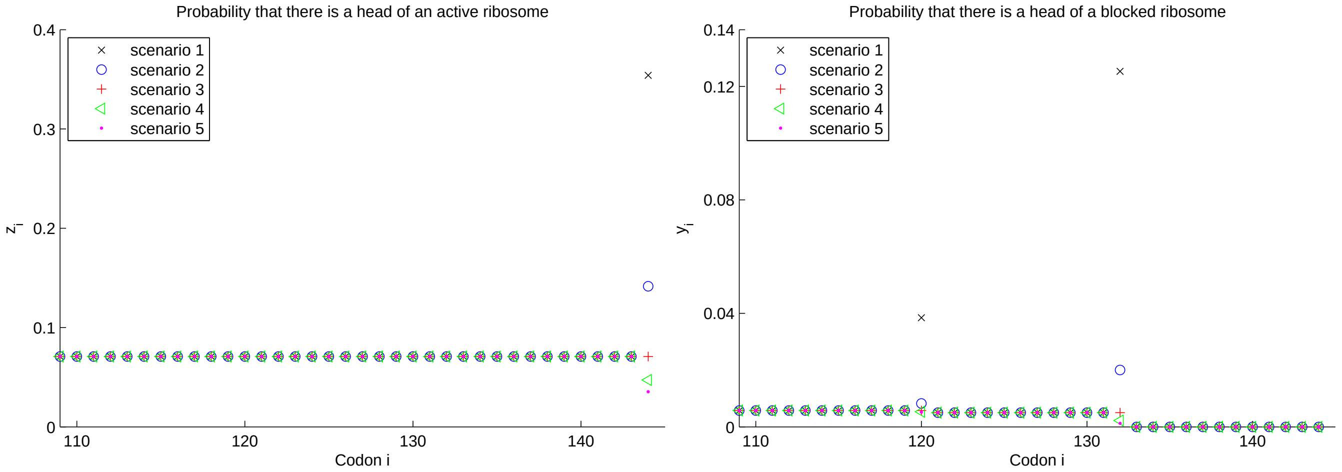

Figure 3: The left (resp. right) plot displays the probability that a codon is occupied by the head of an active (resp. blocked) ribosome for scenarios with increasing termination rates.

5.2 Performance on large-scale networks

We evaluate the scalability of the model for general tandem single server bufferless networks (Systems (14) and (15)). We consider a classical setting where blocking is triggered by the queue directly downstream being full, i.e. $L = 1$.

We solve the Systems of Equations (14) and (15) for all queues simultaneously with the procedure described in Section 5.1. The distributions are initialized with the point: $(t_i, y_i, z_i) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$.

We consider networks with varying number of queues, N. The number of queues varies from 100 to 1000 with a step size of 100 (i.e. 100, 200, 300,...,1000), from 2000 to 10,000 with a step size of 1000, and from 20,000 to 100,000 with a step size of 10,000. That is we consider a total of 28 network sizes varying from 100 to 100,000 queues.

For each network size, we consider a set of four scenarios with varying levels of congestion, i.e. we fix the service rate for all queues and vary the external arrival rates. For all four scenarios all queues have a common service rate equal to 1. The external arrival rates are given in Table 1. These four scenarios have increasing levels of congestion.

For all four scenarios and all network sizes convergence was reached with either 5 or 6 iterations. Since the number of iterations is constant across these scenarios, it is not sensitive to the level of congestion. Furthermore, the number of iterations is also insensitive to the network size. That is, convergence is reached for large-scale networks with few iterations.

We also evaluate the performance of this method, by analyzing the total time until convergence. Figures 4, 5 and 6 display the time until convergence for all four scenarios and 28 network sizes.

Scenario	1	2	3	4
γ	0.5	0.6	0.7	0.8

Table 1: External arrival rate scenarios

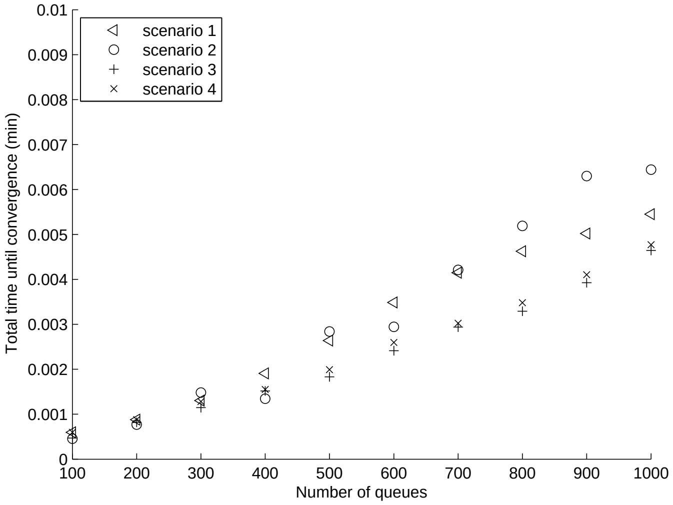

Figure 4: Time until convergence for scenarios 1-4 and small-scale networks

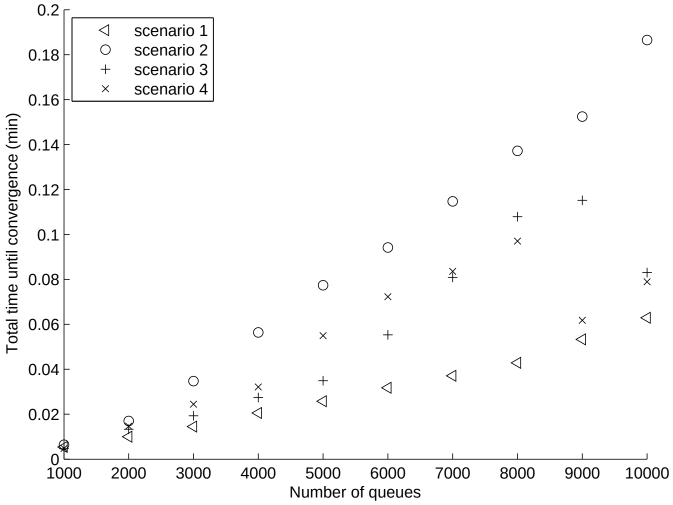

Figure 5: Time until convergence for scenarios 1-4 and medium-scale networks

Figures 4 and 5 indicate that for small and medium-scale networks (with less than $10,000$ queues) and for various congestion levels, the time increases linearly, and remains under 0.2 minutes. For large-scale networks (Figure 6), the time also increases linearly, and is of the order of several minutes.

6 Conclusions

This paper presents an analytical queueing model to evaluate congestion in tandem single server bufferless networks, and in particular in protein synthesis networks. Each codon of an mRNA is modeled as a queue. Each mRNA strand is modeled as a tandem network of single server bufferless queues. The methodology derives a distribution for each codon, that evaluates whether or not there is a head of a ribosome on that codon, and in particular identifies whether ribosomes are blocked by downstream ribosomes. This state space formulation leads to a more detailed quantification of the performance of congested networks. The model generalizes the concept of blocking: blocking events can be triggered by an arbitrary set of queues.

This approach builds upon the model in Osorio and Bierlaire (2009). The model consists of a system of linear and quadratic equations, that can be decoupled. We illustrate the use of this model to evaluate the location of the ribosomes along an mRNA strand, and in particular to quantify ribosome blocking. We evaluate the scalability of this method, by considering networks with varying levels of congestion with up to $100,000$ queues. The method is numerically efficient, and is therefore suitable for large-scale instances.

We are currently working with the Laboratory of Computational Systems Biotechnology to compare the distributional estimates of this approach versus those proposed by other protein synthesis methods, including that of Mehra and Hatzimanikatis (2006). Given the lack of experimental data,

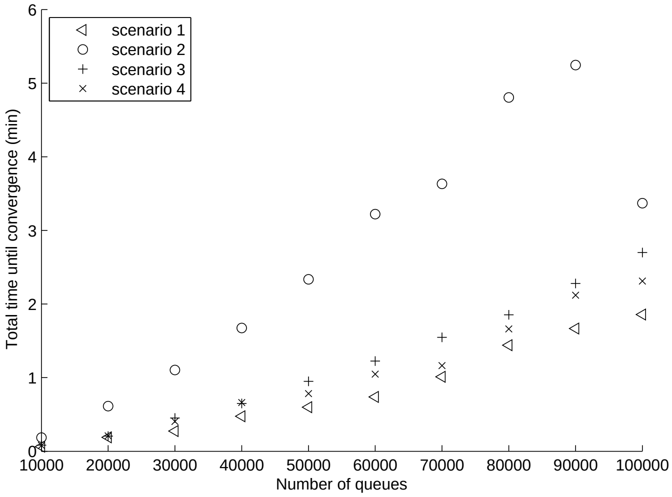

Figure 6: Time until convergence for scenarios 1-4 and large-scale networks

it is intricate to draw conclusions from the differences in these estimates. Nevertheless, it is of interest to investigate their numerical performance and in particular to compare their scalability. Once the validation phase has been completed, this model will be applied to a more general contexts considering multiple mRNA species and codon-specific elongation rates. These factors are captured by the exogenous parameters of our model; taking them into account is therefore straightforward.

Acknowledgments

A project in collaboration with the Laboratory of Computational Systems Biotechnology (LCSB) at Ecole Polytechnique Fédérale de Lausanne directed by Prof. Vassily Hatzimanikatis lead to the formulation of this model. The authors thank the members of LCSB, and in particular Prof. Vassily Hatzimanikatis and Dr. Luis Mier-y-Teran. This research was supported by the Swiss National Science Foundation grant 205320-117581.

Appendix

Derivation of the unblocking rate for an arbitrary blocking structure

We detail how the unblocking rate is derived when allowing for an arbitrary blocking structure. This description follows that of Osorio and Bierlaire (2009) (Section 4.2.3.1 in that paper). The scalar $\tilde{\mu}_i$ denotes the rate at which a trigger queue of queue i unblocks jobs that are blocked at queue i. We denote by s_{ij}, the proportion of jobs blocked by queue j that are blocked at queue i, i.e.

$$s_{ij} = \tilde{q}_{ij}\lambda_i^{\text{eff}}/(\sum_k \tilde{q_{kj}}\lambda_k^{\text{eff}}).$$

Suppose queue j is blocking jobs. It is therefore full and is serving at rate μ_j^{eff}. It unblocks jobs that are blocked at queue i at the rate $s_{ij}\mu_j^{\text{eff}}$. By averaging over the possible trigger queues of queue i we obtain an approximation for $\tilde{\mu}_i$:

$$\frac{1}{\tilde{\mu}_i} = \sum_j \tilde{q}_{ij} \frac{1}{s_{ij}\mu_j^{\text{eff}}} = \sum_{j \in \mathcal{T}_i} \frac{\sum_k \tilde{q_{kj}} \lambda_k^{\text{eff}}}{\lambda_i^{\text{eff}} \mu_j^{\text{eff}}}. \tag{27}$$

Proof of Lemma 1

We first show that $H(N-L)$ holds. Combining Equations (17c) and (10) yields:

$$\mu_{N-L} z_{N-L} = \gamma t_1. \tag{28}$$

Inserting this into (17b) gives:

$$y_{N-L} = (\mathcal{P}_{N-L} \gamma t_1)/\tilde{\mu}_{N-L}. \tag{29}$$

Hereafter, we denote in brackets the equations used at each step. Since queue $N-L$ is non-terminal, System (12) applies:

$$\begin{aligned} y_{N-L} &= \frac{\mathcal{P}_{N-L}\gamma t_1}{\mu_N^{\text{eff}}} && [12c] \\ &= \frac{(1-t_N)\gamma t_1}{\mu_N^{\text{eff}}}. && [12d] \end{aligned} \tag{30}$$

Since queue N is terminal, Systems (11) and (14) apply:

$$\begin{aligned} y_{N-L} &= \frac{(1-t_N)\gamma t_1}{\mu_N} && [11b] \\ &= \frac{z_N \gamma t_1}{\mu_N} && [14a] \\ &= z_N^2 && [14c] \\ &= (y_N + z_N)^2. && [14b] \end{aligned} \tag{31}$$

This gives $H(N-L)$.

We assume that $H(k)$ holds $\forall k \in [i+1, N-L]$. Since queue i is non-terminal, we can proceed as for $H(N-L)$:

$$\begin{aligned} y_i &= \frac{\mathcal{P}_i \mu_i z_i}{\tilde{\mu}_i} && [17b] \\ &= \frac{\mathcal{P}_i \lambda_i t_i}{\tilde{\mu}_i} && [17c] \\ &= \frac{P_i \gamma t_1}{\tilde{\mu}_i} && [10] \\ &= \frac{P_i \gamma t_1}{\mu_{i+L}^{\text{eff}}} && [12c] \\ &= \frac{(1-t_{i+L})\gamma t_1}{\mu_{i+L}^{\text{eff}}}. && [12d] \end{aligned} \tag{32}$$

We distinguish between two cases. Firstly, if queue $i+L$ is terminal we have:

$$\begin{aligned} y_i &= \frac{(1-t_{i+L})\gamma t_1}{\mu_{i+L}} && [11b] \\ &= \frac{z_{i+L}\gamma t_1}{\mu_{i+L}} && [14a] \\ &= (z_{i+L})^2 && [14c] \\ &= (y_{i+L}+z_{i+L})^2. && [14b] \end{aligned} \tag{33}$$

Secondly, if queue $i+L$ is non-terminal, then:

$$\begin{aligned} y_i &= (1-t_{i+L})\gamma t_1 \left(\frac{1}{\mu_{i+L}} + (1-t_{i+2L})\frac{1}{\mu^{\text{eff}}_{i+2L}}\right) && [12b] \\ &= (1-t_{i+L})\gamma t_1 \left(\frac{1}{\mu_{i+L}} + (1-t_{i+2L})\frac{1}{\tilde{\mu}_{i+L}}\right) && [12c] \\ &= (1-t_{i+L})\gamma t_1 \left(\frac{1}{\mu_{i+L}} + \mathcal{P}_{i+L}\frac{1}{\tilde{\mu}_{i+L}}\right) && [12d] \\ &= (1-t_{i+L})\lambda_{i+L} t_{i+L} \left(\frac{1}{\mu_{i+L}} + \mathcal{P}_{i+L}\frac{1}{\tilde{\mu}_{i+L}}\right) && [10] \end{aligned} \tag{34}$$

$$\begin{aligned} y_i &= (1-t_{i+L})\mu_{i+L}\ z_{i+L} \left(\frac{1}{\mu_{i+L}} + \mathcal{P}_{i+L}\frac{1}{\tilde{\mu}_{i+L}}\right) && [17c] \\ &= (y_{i+L}+z_{i+L}) \left(z_{i+L} + \mathcal{P}_{i+L}\ \mu_{i+L}\ z_{i+L}\frac{1}{\tilde{\mu}_{i+L}}\right) && [17a] \\ &= (y_{i+L}+z_{i+L})(z_{i+L}+y_{i+L}) && [17b] \\ &= (y_{i+L}+z_{i+L})^2. \end{aligned} \tag{35}$$

This concludes the recurrence. □

References

Akyildiz, I. F. and von Brand, H. (1994). Exact solutions to networks of queues with blocking-after-service, *Theoret. Comput. Sci.* **125**(1): 111–130.

Alfa, A. S. and Liu, B. (2004). Performance analysis of a mobile communication network: the tandem case, *Comp. Comm.* **27**(3): 208–221.

Altiok, T. (1982). Approximate analysis of exponential tandem queues with blocking, *European Journal of Operational Research* **11**(4): 390–398.

Altiok, T. (1989). Approximate analysis of queues in series with phase-type service times and blocking, *Oper. Res.* **37**(4): 601–610.

Artalejo, J. R. (1999). Accessible bibliography on retrial queues, *Math. Comput. Modelling* **30**(3-4): 1–6.

Balsamo, S., De Nitto Persone, V. and Inverardi, P. (2003). A review on queueing network models with finite capacity queues for software architectures performance prediction, *Perf. Evaluation* **51**(2-4): 269–288.

Balsamo, S., De Nitto Persone, V. and Onvural, R. (2001). *Analysis of Queueing Networks with Blocking*, Vol. 31 of *International Series in Operations Research and Management Science*, Kluwer Academic Publishers, Boston.

Cruz, F. R. B., Smith, J. M. and Queiroz, D. C. (2005). Service and capacity allocation in m/g/c/c state-dependent queueing networks, *Computers & Operations Research* **32**(6): 1545–1563.

Dallery, Y. and Frein, Y. (1993). On decomposition methods for tandem queueing networks with blocking, *Operations Research* **41**(2): 386–399.

Dallery, Y. and Gershwin, S. B. (1992). Manufacturing flow line systems: a review of models and analytical results, *Queueing Systems* **12**(1-2): 3–94.

Dennis, J. E. and Schnabel, R. B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations*, Vol. 16 of *Classics in Applied Mathematics*, SIAM, Philadelphia.

Gelenbe, E. (2007). Steady-state solution of probabilistic gene regulatory networks, *Physical Review E* **76**(3).

Gelenbe, E. (2008). Network of interacting synthetic molecules in steady state, *Proceedings of the Royal Society A* **464**: 2219–2228.

Gershwin, S. B. (1987). An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking, *Operations Research* **35**(2): 291–305.

Grassman, W. and Derkic, S. (2000). An analytical solution for a tandem queue with blocking, *Queueing Syst.* **36**(1-3): 221–235.

Jun, K. P. and Perros, H. G. (1990). An approximate analysis of open tandem queueing networks with blocking and general service times, *European Journal of Operational Research* **46**(1): 123–135.

Kerbache, L. and Smith, J. M. (2000). Multi-objective routing within large scale facilities using open finite queueing networks, *European Journal of Operational Research* **121**(1): 105–123.

Mathworks, Inc. (2008). *Optimization Toolbox Version 4. User's Guide Matlab*, Natick, MA, USA.

Mehra, A. and Hatzimanikatis, V. (2006). An algorithmic framework for genome-wide modeling and analysis of translation networks, *Biophysical Journal* **90**: 1136–1146.

Mier-y-Teran-Romero, L., Silber, M. and Hatzimanikatis, V. (2009). The origins of time-delay in template biopolymerization processes, *Technical report*, Laboratory of Computational Systems Biotechnology, SB, Ecole Polytechnique Fédérale de Lausanne.

Osorio, C. (2010). *Mitigating network congestion: analytical models, optimization methods and their applications*, PhD thesis, Ecole Polytechnique Fédérale de Lausanne.

Osorio, C. and Bierlaire, M. (2009). An analytic finite capacity queueing network model capturing the propagation of congestion and blocking, *European Journal of Operational Research* **196**(3): 996–1007.

Papadopoulos, H. T. and Heavey, C. (1996). Queueing theory in manufacturing systems analysis and design: A classification of models for production and transfer lines, *European Journal of Operational Research* **92**(1): 1–27.

Perros, H. (1984). Queueing networks with blocking: A bibliography, *ACM SIGMETRICS Performance Evaluation Review* **12**(2): 8–12.

Perros, H. (1994). *Queueing networks with blocking: Exact and Approximate Solutions*, Oxford University Press, New York, NY, USA.

Perros, H. (2003). Open queueing networks with blocking - a personal log, *in* G. Kotsis (ed.), *Performance Evaluation - Stories and Perspectives*, Austrian Computer Society, Vienna, Austria, pp. 105–115.

van Vuuren, M., Adan, I. J. B. F. and Resing-Sassen, S. A. E. (2005). Performance analysis of multi-server tandem queues with finite buffers and blocking, *OR Spectrum* **27**(2-3): 315–338.