Approximation of the Transient Joint Queue-Length Distribution in Tandem Markovian Networks

by
Jana H. Yamani

B.S. Computer Science and Mathematics
Northeastern University, 2009

Submitted to the School of Engineering in partial fulfillment of the requirements for the degree of

Master of Science in Computation for Design and Optimization
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September 2013

Author . . . .
The School of Engineering
August 26, 2013

Certified by . . . .
Carolina Osorio
Assistant Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by . . . .
Nicolas Hadjiconstatinou
Professor of Mechanical Engineering
Director, Computation for Design and Optimization (CDO)

Approximation of the Transient Joint Queue-Length Distribution in Tandem Markovian Networks

by
Jana H. Yamani

Submitted to the School of Engineering
on August 26, 2013, in partial fulfillment of the
requirements for the degree of
Master of Science in Computation for Design and Optimization

Abstract

This work considers an urban traffic network, and represents it as a Markovian queueing network. This work proposes an analytical approximation of the time-dependent joint queue-length distribution of the network. The challenge is to provide an accurate analytical description of between and within queue (i.e. link) dynamics, while deriving a tractable approach. In order to achieve this, we use an aggregate description of queue states (i.e. state space reduction). These are referred to as aggregate (queue-length) distributions. This reduces the dimensionality of the joint distribution.

The proposed method is formulated over three different stages: we approximate the time-dependent aggregate distribution of 1) a single queue, 2) a tandem 3-queue network, 3) a tandem network of arbitrary size. The third stage decomposes the network into overlapping 3-queue sub-networks. The methods are validated versus simulation results. We then use the proposed tandem network model to solve an urban traffic signal control problem, and analyze the added value of accounting for time-dependent between queue dependency in traffic management problems for congested urban networks.

Thesis Supervisor: Carolina Osorio
Title: Assistant Professor of Civil and Environmental Engineering

Acknowledgements

This work was completed under the supervision of my research advisor, Professor Carolina Osorio. Carolina, an enormous thank you goes to you for believing in me, for encouraging me to take on new challenges every step of the way and for taking me in late and making it possible for me to complete this thesis in less than a year. I hope to always be in touch.

To my husband, Abdulrahman Tarabzouni: A special thank you goes to you for encouraging me to come to MIT even when you knew you were going to be thousands of miles away from Leen and I. Your constant support, love and encouragement are what helped me overcome obstacles.

To my daughter, Leen Tarabzouni: You are my star. Thank you for reminding me that family is the most important thing, and for forcing me to take breaks from studying! It was great seeing you grow to be this wonderful positive person that you are today. It was great sharing the MIT journey with you. I hope one day you will go through it yourself.

To my parents, siblings and in-laws: Thank you for your support and encouragement in tough times, and for reminding me that having my dreams met is something that I should work so hard for.

To the CDO administrator, Barbara Lechner: Thank you for being there when I needed you. You were always helping me get through academic and personal hardships. I wouldn’t have done it without your constant support. You’ve been extremely missed. May you rest in peace.

To all my fellow friends at CDO: You have made my stay at MIT a memorable one. Thanks to each one of you. I hope to be in touch always.

To my fellowship sponsor, King Abdullah of Saudi Arabia, the Ministry of higher education and the Saudi Arabian Cultural Mission: thank you for your generous help and support. I am lucky to come from a country that supports education in all possible ways.

Contents

Cover page 1

Abstract 3

Acknowledgements 5

Contents 7

List of figures 9

List of tables 11

Chapter 1. Introduction 13

1.1 Literature review 14
1.2 Model background 18
1.3 Overview 19

Chapter 2. Model formulation 21

2.1 Aggregation-disaggregation framework 21
2.2 Aggregate transient model for a single and a network of tandem queues.25
2.2.1 Aggregate transient model for a single queue 25
2.2.2 Aggregate transient model for a three-queue tandem network. . . 33
2.2.3 Aggregate transient model for an M-queue tandem network. 46

Chapter 3. Validation 61

3.1 Single queue. 62
3.2 Three-queue tandem network. 68
3.3 M-queue tandem network. 73
3.3.1 Five queue network 73
3.3.2 Eight queue network. 74
3.3.3 Twenty-five queue network. 77

Chapter 4. Case study 79

4.1 Network. 79
4.2 Problem formulation. 81
4.2 Implementation notes. 84
4.4 Results. 85
4.4.1 Medium demand scenario. 85

4.4.2 High demand scenario. 86

Chapter 5. Conclusion **89**

Appendix **91**

A Transition rate matrix for a three queue network. 91

Bibliography **95**

List of figures

2-1 Aggregating the state space of a single queue to three aggregate states 23
2-2 Decomposing the network to overlapping sub-networks of three queues 46

3-1 Results of experiment 1 at t=1,10,50 and error plots 63
3-2 Results of experiment 2 at t=1,10,50 and error plots .64
3-3 Results of experiment 3 at t=1,10,50 and error plots 64
3-4 Results of experiment 4 at t=1,10, 50 and error plots .65
3-5 Results of experiment 5 at t=1,10, 50 and error plots 65
3-6 Results of experiment 6 at t=1,10, 50and error plots. 66
3-7 Results of experiment 7 at t=1,10, 50 and error plots .66
3-8 Results of experiment 8 at t=1,10, 50and error plots. 67
3-9 Results of experiment 9 at t=1,10, 50 and error plots .67
3-10 Results of experiment 10 at t=1,10, 50and error plots. 68
3-11 Results of experiment 1,2,3 for the 3 joint queue-length distribution. 70
3-12 Results of experiment 4,5,6 for the 3 joint queue-length distribution. 71
3-13 Results of experiment 7,8,9 for the 3 joint queue-length distribution. 72
3-14 Results of the 5-queue tandem network experiment 74
3-15 Results of the 8-queue tandem network experiment at t=1. 75
3-16 Results of the 8-queue tandem network experiment at t=10. 76
3-17 Results of the 8-queue tandem network experiment at t=50. 76
3-18 Histogram of the 25-queue tandem network experiment errors at t=1. 77
3-19 Histogram of the 25-queue tandem network experiment errors at t=10. 78
3-20 Histogram of the 25-queue tandem network experiment errors at t=50. 78

4-1 Urban road network of single roads for the case study. 80
4-2 CDF’s of the average trip time for the medium demand test. 86
4-2 CDF’s of the average trip time for the high demand test. 87

List of tables

2-1 All blocking scenarios with their detailed information. 35

3-1 Experiments to test a single queue 63
3-2 Experiments to test a three-queue network. 69
3-3 Experiment to test a 5 queue tandem network 73
3-4 Experiment to test an 8-queue tandem network. 75

4-1 Demand in vehicle/hour for the medium and high demand scenarios 80

Chapter 1. Introduction

In urban traffic networks, to reduce congestion and improve network-wide performance, one must understand two aspects of the network: the dynamics within each link (i.e., road), and the possibilities of blockings to occur and propagate over time. Blocking occurs when a customer completes service in a link but cannot proceed downstream because the downstream link is full. Queueing theory helps in analyzing both aspects of the network by modeling links as queues. One can study the behavior of queues over time if the arrival process of customers and service mechanism are known. In this thesis, we will represent an urban road network as a Markovian finite capacity queueing network. We are interested in understanding the distribution of customers in the network at any point in time, which can be done through the analysis of the transient joint queue-length distribution (denoted transient joint distribution hereafter) of the network. Calculating the exact transient joint distribution is a computationally expensive task given the high dimensional system of differential equations to be solved; hence, the objective of this thesis is to analytically approximate the transient joint queue-length distribution of the network.

We will specifically look at M/M/1/K queues. The number of customers in an M/M/1/K queue is defined as a stochastic process, its state space is the set {0,1,2…,K}, where K is the state capacity of the queue. This type of queue is governed by independent identically distributed (iid) exponential interarrival times with arrival rate λ and iid exponential service times with service rate μ. M/M/1/K queues are the most elementary of finite capacity queueing models (Strugul, 2000). They are also appealing to study because of the availability of closed-form expressions that describe a wide range of queue metrics.

1.2 Literature review

Calculating the exact transient queue-length distribution of a network requires working with exponentials of high-dimensional matrices that are computationally expensive to compute. Due to the mathematical difficulty of computing the transient distribution of a network, researchers have previously focused on developing models that calculate the steady-state distribution instead of the transient distribution (Phillips, 1995). In cases where there is a need to understand the transient distribution of the network before it reaches to steady-state or when the system does not reach a steady state, the transient solution accurately portrays the behavior of the system as opposed to the stationary which if exists showcases only the final state of the network (Kaczynski, Leemis and Drew, 2012).

Although the literature focuses on steady-state queueing models (Phillips, 1995), transient queueing models have been studied and developed by researchers. In this section, we will focus our investigation on models that look at finite capacity queues and yield expressions for the transient queue-length distributions for a single queue or a network of queues. These models are generally classified into three groups: exact models, analytical approximation models, and numerical approximation models.

The first exact closed-form expression to the transient queue-length distribution of an M/M/1/K queue was developed by Morse (1958, p.65-67). Morse's closed-form equation expresses the transient distribution as the sum of the steady state solution and a transient term. As time increases in the network, the transient term becomes negligible compared to the steady-state solution. The transient solution given by Morse, while useful for a single queue, does not allow us to model a joint queue-length distribution of multiple queues. Takacs (1961) also derived a closed form that yields the same results as Morse (1958) and also has the same limitations.

Another exact model is one developed by Parthasarathy (1987), which derives a transient expression for a single M/M/1/K queues that include integrals of Bessel functions. With small modifications, the expression given can be applied to different types of queues including single or multiple server queues, and queues with or without balking. For instance, Abou-El-Ata (1993) extended Parthasarathy's work to solve the transient behavior of an M/M/1/K queue with balking customers. Despite the fact that the transient expression can be applied to different queue types, the integrals of Bessel functions are complex and hard to accurately compute since they are defined as an infinite series. Given the above, exact models have certain limitations and complexities that can be overcome by approximate models.

When it comes to analytical approximation models, Stern's (1979) model for a single M/M/1 queue uses the form of the queue-length distribution of the exact model in his approximation. The transient queue-length distribution is then expressed as a sum of exponential terms. The expression of the transient is then transformed to a form where the eigenvalues and vectors of the expression is used. Stern shows that the expression for the marginal distribution is in a form that lends itself to simple approximation for the transient mean queue-length. Not only does this method apply for a single queue, a similar approach can be taken to obtain an approximation for the joint distribution of a network of queues. While this model seems to work well for any degree of accuracy, it is crucial to use a small time-step when computing the queue-length distributions, which would result in longer running periods.

Filipiak's (1988) model is another example of an analytical approximation model for calculating the transient queue-length distribution of a single M/M//K queue. The model is called a fluid flow approximation because the core of the model consists of differential equations describing the rate of flow of customers into and out of the queue and relating it to the transient distribution of the queue (Phillips, 1995). The differential equations contain some characteristic functions that if their roots were found, yield the transient distribution for the M/M/1/K queue.

Filipiak's method was then extended by Phillips (1995). Phillips' method however uses different characteristic functions that are easier to solve roots for. Either way, solving roots of high degree polynomials are usually expensive and time-consuming to compute.

Apart from analytical approximation models, numerical approximation models have also been developed to evaluate the transient queue-length distribution. These methods, however, deal directly with the differential equations of the queue-length distributions, which in most cases are high-dimensional systems to solve (Rothkopf and Oren, 1979). Grassmann's paper (1977), for instance, explores three different numerical methods to solve the transient queue-length distribution of M/M/1/K queues. The three methods are: Rung-Kutta, Modified Runge-Kutta and Liou, and Randomization. The methods are closely related, yet the randomization method is shown to be superior than the others. An important trait that these methods exploit is that they preserve the sparsity of the transition rate matrix. It is also important to note that these methods can be applied to solve the queue-length distribution of a single Markovian queue or the joint queue-length distribution of a network of Markovian queues.

Despite the fact that numerical methods have very low execution time compared to exact and analytical approximation methods, the main problem faced by many authors is the high dimensional system of differential equations being solved. A queueing system with *n* queues leads to *n*-tuple states. There is then $\prod_{i=1}^{n} K_i$ different states, where K_i is the capacity of queue i. The transition rate matrix will then be of dimension $(\prod_{i=1}^{n} K_i)^2$. Even for small values of K_i and n, this number can be very large and very hard to store (Grassmann, 1977).

Dealing with a network with large numbers of queues or large queue capacities have been found challenging for many of the methods above. One way to reduce the dimensions of the system of equations being solved is by aggregating the queue-length state space. The aggregation process is done by combining some states into an aggregate states. Aggregation of queue states for stationary Markov chain was introduced by Takahashi (1975). Takahashi later extended the previous work to propose an exact numerical derivation of a marginal aggregate queue-length distribution and a joint aggregate queue-length distribution (Takahashi and Song, 1991). In Takahashi and Song's paper (1991), they enhanced the aggregation model by modeling the joint queue-length distribution of adjacent queues, therefore accounting for any blockings between queues. They showed an example of approximating the stationary distribution for a 5-queue tandem network with blocking by looking at joints with different number of queues. They first looked at individual queues in the network and calculated the marginal queue-length distribution of each queue independently. They then looked at two queues at a time and calculated the two-queue joint queue-length distribution. Lastly, they looked at three queues and higher at a time and calculated the three-queue or more joint queue-length distribution. They showed that the higher the number of queues represented in the joint, the more accurate the stationary distribution is. The reason is because calculating joint distributions with more queues means accounting for more between-queue activities including blockings (Takahashi and Song, 1991).

The papers on aggregation-disaggregation from Takahashi tackled two of the challenges of estimating the stationary queue-length distribution: the size of the system and the dependencies between queues that lead to blocking. The work done by Takahashi was then extended by Schweitzer (1984) to introduce the same aggregation-disaggregation techniques for the transient analysis of Markov chains and it's application to queueing networks. Schweitzer's approach tackles the same transient model challenges, but also ensures the convergence to stationary distribution.

Most of the work in this thesis combines ideas from both exact and analytical approximation models surveyed above, as well as aggregation-disaggregation techniques from Takahashi and Schweitzer.

1.2 Model background

To introduce the model, we introduce the following notation:

$X(t)$	number of customers in the queue at time *t*;
K	queue capacity;
Ω	state space of the Markovian queue;
Q	transition rate matrix for a single queue;
q_{ij}	transition rate from state *i* to state *j*;
λ	customer arrival rate to the queue;
μ	service rate of the queue;
$p_i(t)$	probability of being in state *i* at time *t;*
$p(t)$	row vector representing the transient queue-length distribution of a queue;
p^0	initial queue-length distribution.

Let $\{X(t), t \geq 0\}$ represent a finite-state continuous-time Markovian queueing system with state space *Ω* and state space dimension K+1, where the states represent the number of customers in the system. For a single queue, the transition rate matrix is given by $Q = [q_{ij}]$, with values $q_{i(i+1)} = \lambda,\ q_{i(i-1)} = \mu$. The diagonal elements are given by,

$$q_{ii} = \sum_{j=1, j\neq i}^{K+1} -q_{ij}, \tag{1}$$

and all other terms being null.

Let $p_i(t)$ be the probability that the queue has *i* customers at time *t*, then the row vector $p(t)$ represents the transient queue-length distribution of all states. The behavior of the finite Markovian queue can be described by the Kolmogorov system of differential equations (Muppala and Trivedi, 1992):

$$\frac{d}{dt}p(t) = p(t)Q, \qquad p(0) = p_0. \tag{2}$$

Here, p_0 represents the initial queue-length distribution of the Markovian queue. The solution of this system of first order linear differential equations yields the transient queue-length distribution of the queue at time *t,* $p(t)$. Several methods for solving the differential equations are available. For instance, differential equation solver like Runge-Kutta or Randomization (Grassmann, 1977) can solve this numerically. However, we are interested in solving this equation analytically.

We can write the general solution of equation (2) as:

$$p(t) = p(0)e^{Qt} = p^0 \, e^{\mathrm{Qt}}. \tag{3}$$

We can rewrite equation (3) differently, by shifting the origin of the time axis to t_1 instead of 0 since the process is time-homogeneous (Grassman, 1977):

$$p(t_2) = p(t_1)e^{Q(t_2-t_1)}. \tag{4}$$

For a single queue, it is convenient to solve the transient queue-length distribution using equations (3) or (4). However, the dimensions of Q increases exponentially as the number of queues in the network or capacities of the queues get larger. In addition, direct evaluation of the matrix exponential can run into high accumulation of round-off errors since the Q matrix contains both positive and negative entries. In the next chapter we will present a model that accounts for these challenges.

1.3 Overview

The remainder of this thesis is structured as follows.

In chapter 2, we will formulate the model. We will present the aggregation-disaggregation framework, and then apply the aggregation on a single queue, a 3-queue tandem network, and an M-queue tandem network. We will present the analytical approximation model of the transient queue-length distribution of an M-queue tandem network in the last section of the chapter.

In chapter 3, we will validate the model by comparing the transient joint distribution obtained from our model against those estimated from an exact model for one queue and a discrete event simulation model for a network of queues.

In chapter 4, we will apply the transient model to a traditional signal control problem on a network to measure the added value of accounting for the transient behavior. We will evaluate multiple scenarios that consider the same road network and different travel demands. Our interest is to see how our model performs with different demand scenarios compared to a stationary joint model.

Finally, in chapter 5, we will present a summary of the model and of the results from the case study, and show the added value for accounting for the transient joint distribution.

Chapter 2. Model formulation

2.1 Aggregation-disaggregation framework

For us to overcome the dimensionality problem mentioned in the first chapter, we apply Schweitzer's (1984) aggregation technique for transient Markovian queueing systems. The technique assumes a finite-state Markovian queueing system with aperiodic and communicative properties. The urban transportation network that we're looking to analyze meets all the assumption addressed by Schweitzer.

To present the framework, we first introduce the following notation:

Ω state space of the Markovian queueing system;

M size of Ω;

$\bar{\Omega}$ aggregate state space of the Markovian queueing system;

Ω_a state space representing all disaggregate states that are in aggregate state *a;*

$\bar{M}$ size of $\bar{\Omega}$;

N disaggregate state;

A aggregate state;

$p_{N=n}(t)$ probability of being in disaggregate state *n* at time t;

$p_{A=a}(t)$ probability of being in aggregate state *a* at time t;

$p_N(t)$ row vector representing the disaggregate transient queue-length distribution of a queue;

$p_A(t)$ row vector representing the aggregate transient queue-length distribution of a queue;

$\bar{q}_{ab}$ transition rate from aggregate state a$\in \bar{\Omega}$ to aggregate state b $\in \bar{\Omega}$;

$\hat{q}_{aj}(t)$ transition rate from aggregate state a$\in \bar{\Omega}$ to disaggregate state $j \in \Omega$;

$\bar{\lambda}\,(t)$ aggregate arrival rate at time t;

$\bar{\mu}\,(t)$ aggregate service rate at time t.

Assume our Markovian queueing system has a state space Ω of dimension M, the probability of being in any state n ∈ Ω at time t is denoted by $p_{N=n}(t)$, and the transition rate from going from state $i \in \Omega$ to state $j \in \Omega$ is denoted by q_{ij}. To aggregate the state space, we cluster states together to get an aggregated state space $\overline{\Omega}$, of size $\overline{M} < M$. For an aggregate state a ∈ $\overline{\Omega}$, the set Ω_a represents all disaggregate states that are in a. Hence, the probability of being in an aggregate state a denoted $p_{A=a}(t)$, is defined as a function of the disaggregate probabilities,

$$p_{A=a}(t) = \sum_{n \in \Omega_\mathrm{a}} p_{N=n}(t). \tag{5}$$

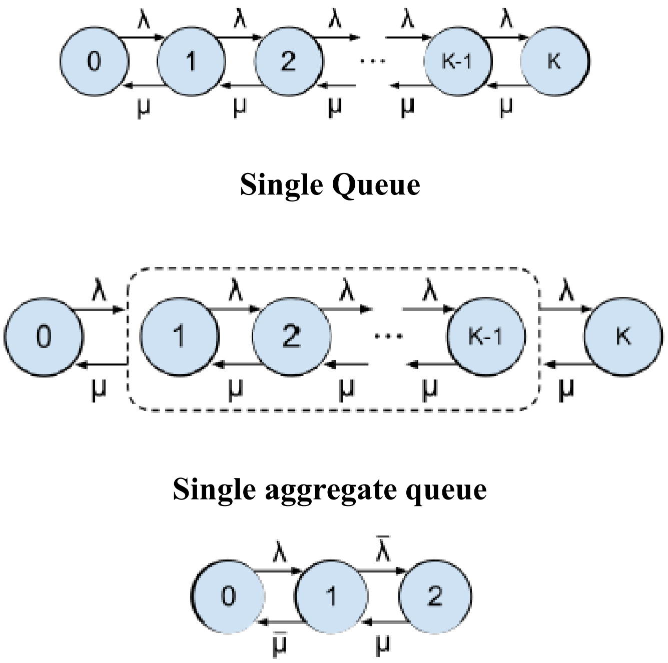
The transition rate $\bar{q}_{ab}$ from aggregate state a∈ $\overline{\Omega}$ to aggregate state b ∈ $\overline{\Omega}$ as defined by Schweitzer (1985) is:

$$\bar{q}_{ab}(t) = \frac{\sum_{j \in \Omega_\mathrm{a}} \sum_{k \in \Omega_\mathrm{b}} p_{N=j}(t)\, q_{jk}(t)}{\sum_{m \in \Omega_\mathrm{a}} p_{N=m}(t)}. \tag{6}$$

Additionally, the transition rate $\hat{q}_{aj}$ from aggregate state a∈ $\overline{\Omega}$ to disaggregate state $j \in \Omega$ as defined by Schweitzer (1985) is:

$$\hat{q}_{aj}(t) = \frac{\sum_{i \in \Omega_\mathrm{a}} p_{A=i}(t)\, q_{ij}(t)}{\sum_{m \in \Omega_\mathrm{a}} p_{N=m}(t)}. \tag{7}$$

In this paper, we use the same decomposition of aggregate states as in Osorio and Wang (2012). Figure 2-1 shows the state transition diagram, before and after aggregating the state space. Each circle in the diagram represents a state, and each arrow represents possible transitions between the states with their rates. Arrivals in the figure are determined by the arrival rate $\lambda \geq 0$, and departures by the service rate $\mu > 0$.



Single Queue

Single aggregate queue

Simplified single aggregate queue

Figure 2-1: Aggregating the state space of a single queue to three aggregate states (Osorio and Wang, 2012)

Initially, we have M= $\mathrm{K} + 1$ states, where K is the queue capacity. We aggregate to get $\bar{\mathrm{M}} = 3$ aggregate states. Our system now has only 3 aggregate states: aggregate state 0 representing an empty queue, aggregate state 2 representing a full queue, and aggregate state 1 representing a non-empty and non-full queue. For a network of queues, this means that the number of equations for the network is linear in the number of queues instead of exponential.

The third image in Figure 2-1 shows that the rates for leaving aggregate state 1 have changed. The other transition rates remain the same because aggregate state 0 and disaggregate state 0 are equivalent. Additionally, aggregate state 2 and disaggregate state K are equivalent. The aggregate system is now fully described by a set of four rates $\lambda, \mu, \bar{\lambda}$ and $\bar{\mu}$. The first two are known and the last two (denoted aggregate arrival rate and aggregate service rate respectively) can be defined from Equations (6), (7) (Osorio and Wang , 2012) and (Schweitzer, 1984) as:

$$\bar{\lambda}\,(t) = \lambda \frac{p_{N=K-1}(t)}{p_{A=1}(t)} = \lambda\, p_{(N=k-1|A=1)}(t)\,, \tag{8}$$

$$\bar{\mu}\,(t) = \mu \frac{p_{N=1}(t)}{p_{A=1}(t)} = \mu\, p_{(N=1|A=1)}(t)\,, \tag{9}$$

where $p_{N=K-1}, p_{N=1}$ are the probabilities that the queue is in disaggregate states K-1, 1 respectively, while $p_{A=1}$ is the probability that the queue is in aggregate state 1.

2.2 Aggregate transient model for a single and a network of tandem queues

In this section, we will apply the aggregation-disaggregation techniques from 2.1 to derive the model for calculating the transient queue-length distribution of a single M/M/1/K queue and the transient joint queue-length distribution for a network of M/M/1/K queues in tandem. We propose to calculate the joint transient distribution of a network of queues in tandem by decomposing the system into overlapping sub-networks of three queues. Below we present this formulation at three different network size levels: a single queue, a network of 3 queues in tandem, and a network of M queues in tandem.

2.2.1 Aggregate transient model for a single queue

For a single finite-capacity Markovian queue, the state space is given by $\Omega = \{0, 1, .., K\}$, where K$\geq$ 0 is the queue capacity. To derive the aggregate model for a single queue-length distribution over time, we will use the same framework introduced in 2.1, where our system now has only 3 aggregate states. This results in a 3x3 aggregate transition rate matrix, Q_A.

The model is implemented in discrete time, and within each time interval, we assume aggregate transition rates to be constant. To present the model, we introduce the following notation:

λ queue arrival rate;

μ queue service rate;

ρ queue traffic intensity;

K queue capacity;

p^0 initial disaggregate queue-length distribution of the queue;

$p_A^k(t)$ aggregate transient queue-length distribution at continuous time t within time interval k;

$p_N^k(t)$	disaggregate transient queue-length distribution at continuous time t within time interval k;
Q_A^k	aggregate transition rate matrix during time interval k;
λ^k	approximated queue arrival rate during time interval *k*;
μ^k	approximated queue service rate during time interval *k*;
ρ^k	approximated queue traffic intensity during time interval k;
$\bar{\lambda}^k$	aggregate arrival rate during time interval k;
$\bar{\mu}^k$	aggregate service rate during time interval k;
P_n	probability of being in disaggregate state *n* at stationarity;
δ	time step length;
T	duration of entire time horizon;
t	continuous time within the $[0, \delta]$ interval.

For a queue with arrival rate λ, service rate μ, capacity K and initial disaggregate queue-length distribution p^0, the traffic intensity ρ is defined as the ratio of the arrival rate to service rate $\rho = \frac{\lambda}{\mu}$. The discrete form of the aggregate queue-length distribution over time is defined as:

$$p_A^k(t) = p_A^{k-1}(\delta) e^{t\,Q_A^k}, \quad \forall t \in [0, \delta], \qquad p_A^k(0) = p_A^{k-1}(\delta), \tag{10a}$$

$$Q_A^k = \begin{bmatrix} -\lambda & \lambda & 0 \\ \bar{\mu}^k & -(\bar{\lambda}^k + \bar{\mu}^k) & \bar{\lambda}^k \\ 0 & \mu & -\mu \end{bmatrix}, \tag{10b}$$

where the initial aggregate queue-length distribution and initial service and arrival rates are defined as:

$$p_A^0 = \begin{bmatrix} p_{N=0}^0 \\ 1 - p_{N=0}^0 - p_{N=K}^0 \\ p_{N=K}^0 \end{bmatrix}, \qquad \bar{\lambda}^1 = \lambda \frac{p_{N=K-1}^0}{p_{A=1}^0}, \qquad \bar{\mu}^1 = \mu \frac{p_{N=1}^0}{p_{A=1}^0}.$$

(10c)

To calculate the aggregate transition rates $\bar{\lambda}^k, \bar{\mu}^k$, we refer to equations (8) and (9). In discrete time, we get:

$$\bar{\lambda}^k = \lambda \frac{p^{k-1}_{N=K-1}(\delta)}{p^{k-1}_{A=1}(\delta)} = \lambda\, p^{k-1}_{(N=k-1|A=1)}(\delta),$$

(11a)

$$\bar{\mu}^k = \mu \frac{p^{k-1}_{N=1}(\delta)}{p^{k-1}_{A=1}(\delta)} = \mu\, p^{k-1}_{(N=1|A=1)}(\delta).$$

(11b)

Equations (11a) and (11b) require calculations of the disaggregate queue-length distributions (i.e., $p^{k-1}_{N=1}(\delta)$, and $p^{k-1}_{N=K-1}(\delta)$). Since these are not available, we apply the closed form expression of the queue-length distribution from Morse's exact method (1958, p.65-67) to approximate the disaggregate distributions. The transient queue-length distribution as derived by Morse (1958) in continuous time is given by:

$$p_{N=n}(T) = \sum_{m=0}^{K} p^0_{N=m}\, d^m_{N=n}(T),$$

$(12a - 1)$

In discrete time at time interval k, the transient queue-length distribution is defined as:

$$p^k_{N=n}(t) = \sum_{m=0}^{K} p^{k-1}_{N=m}(\delta)\, d^{m,k}_{N=n}(t)\,,\ \forall t \in [0, \delta],$$

$(12a - 2)$

where $p^0_{N=m}$ is the initial probability of being in disaggregate state *m*, $p^{k-1}_{N=m}(0)$ is the probability of being in disaggregate state *m* from the previous time step. In continuous time, $d^m_{N=n}(T)$ is defined as:

$$d_{N=n}^{m}(T) = P_n + \frac{2\rho^{\frac{1}{2}(n-m)}}{K+1}\sum_{s=1}^{K}\frac{1}{x_s}\left[\sin\frac{sm\pi}{K+1} - \sqrt{\rho}\sin\frac{s(m+1)\pi}{K+1}\right]\left[\sin\frac{sn\pi}{K+1} - \sqrt{\rho}\sin\frac{s(n+1)\pi}{K+1}\right]e^{-x_s\,\mu\,T},$$

(12b-1)

and in discrete time during time interval k with continuous time t, $d_{N=n}^{m,k}(t)$ is defined as:

$$d_{N=n}^{m,k}(t) = P_n + \frac{2\rho^{\frac{1}{2}(n-m)}}{K+1}\sum_{s=1}^{K}\frac{1}{x_s}\left[\sin\frac{sm\pi}{K+1} - \sqrt{\rho}\sin\frac{s(m+1)\pi}{K+1}\right]\left[\sin\frac{sn\pi}{K+1} - \sqrt{\rho}\sin\frac{s(n+1)\pi}{K+1}\right]e^{-x_s\,\mu\,t},$$

(12b-2)

$$x_s = \frac{\gamma_s}{\mu} = \frac{\lambda + \mu - 2\sqrt{\lambda\mu}\cos\left(\frac{s\pi}{K+1}\right)}{\mu},$$

(12c)

where, $P_n = \frac{(1-\rho)\rho^n}{1-\rho^{K+1}}$ is the stationary distribution of an M/M/1/K queue, and $n \in [0,1, \dots.., K]$. Both n and $k+1$ are exponents in the stationary distribution equation.

To approximate the disaggregate probabilities, $p_{N=1}^{k-1}(\delta)$, and $p_{N=K-1}^{k-1}(\delta)$, we solve a nonlinear system of equations for μ^{k-1}, λ^{k-1}. The nonlinear system consists of two equations: The first states that $p_{N=0}^{k-1}(\delta)$ and $p_{A=0}^{k-1}(\delta)$ are equal, and the second states that $p_{N=K}^{k-1}(\delta)$ and $p_{A=2}^{k-1}(\delta)$ are equal. We end up solving two nonlinear equations for two unknowns. The nonlinear system is defined below in Equations (13) and (14) and in more details in Equations (15) and (16).

$$p_{A=0}^{k-1}(\delta) - \left(\sum_{m=0}^{K} p_{N=m}^{k-2}(\delta)\ d_{N=0}^{m,k-1}(\delta) \right) = 0, \tag{13}$$

$$p_{A=2}^{k-1}(\delta) - \left(\sum_{m=0}^{K} p_{N=m}^{k-2}(\delta)\ \ d_{N=K}^{m,k-1}(\delta) \right) = 0. \tag{14}$$

We plug Equation (12) into (13) and (14) and get

$$\begin{aligned} & p_{A=0}^{k-1}(\delta) \\ & - \left(\sum_{m=0}^{K} p_{N=m}^{k-2}(\delta) \left(P_0 \right.\right. \\ & + \frac{2(\rho^{k-1})^{\frac{-m}{2}}}{K+1} \sum_{s=1}^{K} \frac{\mu^{k-1}}{\lambda^{k-1} + \mu^{k-1} - 2\sqrt{\lambda^{k-1}\mu^{k-1}} \cos\left(\frac{s\pi}{K+1}\right)} \left[\sin\frac{sm\pi}{K+1} \right. \\ & \left.\left.\left. - \sqrt{\rho^{k-1}} \sin\frac{s(m+1)\pi}{K+1} \right] \left[-\sqrt{\rho^{k-1}} \sin\frac{s\pi}{K+1}\ \right] e^{-\left(\lambda^{k-1}+\mu^{k-1}-2\sqrt{\lambda^{k-1}\mu^{k-1}}\cos\left(\frac{s\pi}{K+1}\right)\right)\delta} \right)\right) \\ & = 0, \end{aligned} \tag{15}$$

$$
\begin{aligned}
&p_{A=2}^{k-1}(\delta) \\
&-\left(\sum_{m=0}^{K} p_{N=m}^{k-2}(\delta)\left(P_K\right.\right. \\
&+\frac{2(\rho^{k-1})^{\frac{1}{2}(K-m)}}{K+1}\sum_{s=1}^{K}\frac{\mu^{k-1}}{\lambda^{k-1}+\mu^{k-1}-2\sqrt{\lambda^{k-1}\mu^{k-1}}\cos\left(\frac{s\pi}{K+1}\right)}\left[\sin\frac{sm\pi}{K+1}\right. \\
&\left.-\sqrt{\rho^{k-1}}\sin\frac{s(m+1)\pi}{K+1}\right]\left[\sin\frac{sK\pi}{K+1}\right]e^{-\left(\lambda^{k-1}+\mu^{k-1}-2\sqrt{\lambda^{k-1}\mu^{k-1}}\cos\left(\frac{s\pi}{K+1}\right)\right)\delta}\left.\left.\vphantom{\sum_{m=0}^{K}}\right)\right) \\
&=0,
\end{aligned}
\tag{16}
$$

where λ^{k-1}, μ^{k-1}, are the queue arrival rate and service rate during time interval $k-1$ that we want to solve for, and $\rho^{k-1} = \frac{\lambda^{k-1}}{\mu^{k-1}}$, where $k-1$ represents the time interval index.

Once we solve for λ^{k-1} and μ^{k-1}, we plug them into the discrete form of Equation (12) with the initial disaggregate distribution $p_N^{k-2}(\delta)$ to get the disaggregate probabilities $p_{N=1}^{k-1}(\delta)$, and $p_{N=K-1}^{k-1}(\delta)$. The disaggregate probabilities will then be plugged into Equations (11a) and (11b) to calculate the disaggregate transition rates $\bar{\lambda}^k, \bar{\mu}^k$.

The full algorithm for solving the transient distribution of a single queue can be described in the following steps:

Input:

Arrival rate to the queue: λ

Service rate of the queue: μ

Queue capacity: K

Initial disaggregate queue-length distribution of the queue: p_N^0

Duration of entire time horizon;T

Output :

Assuming that $\frac{T}{\delta}$ is an integer, the output is an approximation of the aggregate queue-length distribution of a queue at time T (in discrete time at time interval $k = \frac{T}{\delta}$): $p_A^{\frac{T}{\delta}}(t)$, where t can be any value between $[0, \delta]$.

Algorithm:

δ can be initiated as any small number.

For $k = 0,1,2, \dots, \frac{T}{\delta}$

If $k = 0$

1) Calculate the initial aggregate distribution (p_A^0) from the initial disaggregate distribution (p_N^0) using the following equation:

$$p_A^0 = \begin{bmatrix} p_{N=0}^0 \\ 1 - p_{N=0}^0 - p_{N=K}^0 \\ p_{N=K}^0 \end{bmatrix},$$

2) Calculate the initial aggregate transition rates $\bar{\lambda}^1, \bar{\mu}^1$:

$$\bar{\lambda}^1 = \lambda \frac{p_{N=K-1}^0}{p_{A=1}^0}, \qquad \bar{\mu}^1 = \mu \frac{p_{N=1}^0}{p_{A=1}^0}.$$

Else

1) The aggregate queue-length distribution for time step k of continuous time t is defined as:

$$p_A^k(t) = p_A^{k-1}(\delta) e^{\delta\, Q_A^k},$$

where $p_A^k(0) = p_A^{k-1}(\delta)$, $Q_A^k = \begin{bmatrix} -\lambda & \lambda & 0 \\ \bar{\mu}^k & -(\bar{\lambda}^k + \bar{\mu}^k) & \bar{\lambda}^k \\ 0 & \mu & -\mu \end{bmatrix}$.

```
        2) Solve the following nonlinear system of equations to obtain $\lambda^k, \mu^k$,
           where $d_{N=0}^{m,k}(\delta)$ is given by Equation (12b-2) :
              $p_{A=0}^{k}(\delta) - \left( \sum_{m=0}^{K} p_{N=m}^{k-1}(\delta)\ d_{N=0}^{m,k}(\delta) \right) = 0,$
              $p_{A=2}^{k}(\delta) - \left( \sum_{m=0}^{K} p_{N=m}^{k-1}(\delta)\ \ d_{N=K}^{m,k-1}(\delta) \right) = 0.$

        3) Plug $\lambda^k, \mu^k$ into Equation (12) to get the disaggregate probabilities of
           being in disaggregate states $1, K-1$:
              $p_{N=K-1}^{k}(\delta) = \sum_{m=0}^{K} p_{N=m}^{k-1}(\delta) d_{N=K-1}^{m,k}(\delta),$
              $p_{N=1}^{k}(\delta) = \sum_{m=0}^{K} p_{N=m}^{k-1}(\delta)\ d_{N=1}^{m,k}(\delta).$

        4) Calculate $\bar{\lambda}^{k+1}, \bar{\mu}^{k+1}$ for the next time step from the following
           equations:
              $\bar{\lambda}^{k+1} = \lambda \frac{p_{N=K-1}^{k}(\delta)}{p_{A=1}^{k}(\delta)},$
              $\bar{\mu}^{k+1} = \mu \frac{p_{N=1}^{k}(\delta)}{p_{A=1}^{k}(\delta)}.$

    End
End
```

2.2.2 Aggregate transient model for a three-queue tandem network

In this section, we consider three M/M/1/K queues in tandem. For this type of network, we want to approximate the aggregate joint queue-length distribution $p^k_{A=(i,j,l)}(t)$ which is defined as the probability that the first, second and third queue, are in aggregate states i, j, l respectively at continuous time $t \in [0, \delta]$ within time interval k. The aggregate state space is defined as the triplets with 27 unique states where $(i, j, l) \in \{0,1,2\}^3$. Therefore, the dimension of the transition rate matrix is independent of the individual queue capacities and is always 27x27.

We introduce the following notation:

N_i disaggregate state of queue *i*;

A_i aggregate state of queue *i*;

γ_i external arrival rate to queue *i*;

μ_i service rate of queue *i*;

K_i capacity of queue *i*;

K_j capacity of the queue corresponding to blocking scenario *j*;

λ^k_j approximated queue arrival rate for blocking scenario *j* during time interval *k*;

μ^k_j approximated queue service rate for blocking scenario *j* during time interval *k*;

ρ^k_j approximated queue traffic intensity for blocking scenario *j* during time interval k;

$p^k_{A=(i,j,l)}(t)$ aggregate joint queue-length distribution at continuous time t within time interval k;

$p^k_{N=(i,j,l)}(t)$ disaggregate joint queue-length distribution at continuous time t within time interval k;

$p^k_{(A_i=a)}(t)$ the marginal probability that queue *i* is in aggregate state *a* at continuous time *t* within time interval k;

$p^k_{(N_i=n)}(t)$ the marginal probability that queue *i* is in disaggregate state *n* at continuous time *t* within time interval k;

$p_{N_i}^0$ initial disaggregate queue-length distribution for queue *i*;

$p_{A_i}^0$ initial aggregate queue-length distribution for queue *i*;

$p_{A=(i,j,l)}^0$ initial aggregate joint queue-length distribution;

$p_{N=(i,j,l)}^0$ initial disaggregate joint queue-length distribution;

Q_{AJ}^k aggregate joint transition rate matrix (*AJ* is a shorthand for aggregate joint) within time interval *k*;

${\alpha_j^e}^k$ empty aggregate transition rate probability for blocking scenario *j* during time interval k;

${\alpha_j^f}^k$ full aggregate transition rate probability for blocking scenario *j* during time interval $k;$

δ time step length;

T duration of entire time horizon;

t continuous time within the $[0, \delta]$ interval.

Each of the three queues in the network has an external arrival rate γ_i, service rate μ_i, queue capacity K_i and initial disaggregate queue-length distribution $p_{N_i}^0$, where *i* ∈ $\{1,2,3\}$. We calculate the initial disaggregate joint distribution $p_{N=(i,j,l)}^0$ by assuming a product-form joint queue-length distribution, i.e., the initial joint can be decomposed as a product of its marginals. Unfortunately, finite-capacity queueing systems, in general, do not have a product-form joint queue-length distribution. The reason for that is because finite-capacity queueing system give rise to blocking which might cause intricate dependency between queues, where service and arrival rates of queues might increase of decrease depending on any blocking that might occur in the system.

When a queue is causing blocking on upstream queues, the service rates of upstream queues might get decreased because of the blocking. Additionally, when the queue causing the blocking has a service completion, service rates of some blocked upstream queues might increase. Hence, calculating the joint is a challenge in that blocking should be captured in all its scenarios and accounting for these dependencies between queues is necessary.

In a three-queue tandem network, where q_1 is the most upstream, q_1 can be either be not blocked or blocked by either q_2 or q_3, and q_2 can either be not blocked or blocked by q_3, and q_3 is always not blocked. This gives us a total of 6 blocking scenarios. The probability of a job being blocked for each of these scenarios has been approximated in Osorio and Wang (2012). Table 2-1 shows all these scenarios with an approximation of their probabilities of occurrence.

Blocking scenario	Joint State	Blocking Probability	Aggregate transition rate probabilities
q_1 not blocked	{(0,1,2), (0,1),(0,1,2)}	0	α_1^e, α_1^f
$q_1$blocked by q_2	{(0,1,2),2,(0,1)}	$B_1 = \frac{\mu_1}{\mu_1 + \mu_2}$	α_2^e, α_2^f
$q_1$blocked by q_3	{(0,1,2),2,2}	$B_2 = \frac{\mu_1}{\mu_1 + \mu_2 + \mu_3} \frac{\mu_2}{\mu_2 + \mu_3} + \frac{\mu_2}{\mu_1 + \mu_2 + \mu_3} \frac{\mu_1}{\mu_1 + \mu_3}$	α_3^e, α_3^f
q_2 not blocked	{(0,1,2), (0,1,2) ,(0,1)}	0	α_4^e, α_4^f

$q_2$blocked by q_3	{(0,1,2), 1 ,2}
	{0,2,2}
		$B_3 = \frac{\mu_2}{\mu_3 + \mu_2}$
		$B_4 = \frac{\mu_2}{\mu_1 + \mu_2 + \mu_3} \frac{\mu_1}{\mu_1 + \mu_3}$
			α_5^e, α_5^f

q_3 not blocked	All states	0	α_6^e, α_6^f

Table 2-1: All blocking scenarios with joint states, blocking probabilities, and aggregate transition rate probabilities

To calculate the transient joint queue-length distribution, we refer to Equation (10) from the single queue model and modify it to apply for the 3-queue joint model. The joint model is also implemented in discrete time, and within each time interval, we assume aggregate transition rates for all blocking scenarios to be constant. The main equations for the joint transient models is presented below in Equations (17) and (18):

$$p^{k}_{A=(i,j,l)}(t) = p^{k-1}_{A=(i,j,l)}(\delta)e^{t\,Q^{k}_{AJ}}, \quad \forall t \in [0,\delta], \qquad p^{k}_{A=(i,j,l)}(0) = p^{k-1}_{A=(i,j,l)}(\delta), \tag{17}$$

where $Q^{k}_{AJ} = f(\gamma, \mu, \alpha^{k}, B)$ is a 27x27 sparse matrix with nonzero elements given in appendix A. The parameters for the matrix are: $\gamma = [\gamma_1, \gamma_2, \gamma_3]$, $\mu = [\mu_1, \mu_2, \mu_3]$, $B = [B_1, B_2, B_3, B_4]$, $\alpha^{k} = [\alpha^{e^{k}}_{1}, \alpha^{f^{k}}_{1}, \alpha^{e^{k}}_{2}, \alpha^{f^{k}}_{2}, \alpha^{e^{k}}_{3}, \alpha^{f^{k}}_{3}, \alpha^{e^{k}}_{4}, \alpha^{f^{k}}_{4}, \alpha^{e^{k}}_{5}, \alpha^{f^{k}}_{5}, \alpha^{e^{k}}_{6}, \alpha^{f^{k}}_{6}]$. The initial aggregate joint queue-length distribution $p^{0}_{A=(i,j,l)}$, is calculated assuming independent initial marginal aggregate queue-length distributions of the three queues. To calculate it, we perform a cross product of the three initial aggregate queue-length distributions, where $p^{0}_{A=(i,j,l)} = p^{0}_{A_1=i}\, p^{0}_{A_2=j} p^{0}_{A_3=l}$.

We define the aggregate transition rate probabilities as follows:

$$\alpha^{e^{k}}_{1} = p^{k-1}_{((N_1=1|\, A_2\neq 2)|(A_1=1|\, A_2\neq 2))}(\delta) = \frac{p^{\mathrm{k-1}}_{(N_1=1\,|A_2\neq 2)}(\delta)}{p^{\mathrm{k-1}}_{(A_1=1|A_2\neq 2)}(\delta)}$$

$$\alpha^{f^{k}}_{1} = p^{\mathrm{k-1}}_{((N_1=K_1-1|\, A_2\neq 2)|(A_1=1|\, A_2\neq 2))}(\delta) = \frac{p^{\mathrm{k-1}}_{(N_1=K_1-1\,|A_2\neq 2)}(\delta)}{p^{\mathrm{k-1}}_{(A_1=1|A_2\neq 2)}(\delta)}$$

$$\alpha^{e^{k}}_{2} = p^{\mathrm{k-1}}_{((N_1=1|\, A_2=2,A_3\neq 2)|\, (A_1=1|\, A_2=2,A_3\neq 2))}(\delta) = \frac{p^{\mathrm{k-1}}_{(N_1=1\,|A_2=2,A_3\neq 2)}(\delta)}{p^{\mathrm{k-1}}_{(A_1=1|\, A_2=2,A_3\neq 2)}(\delta)}$$

$$\alpha^{f^{k}}_{2} = p^{\mathrm{k-1}}_{((N_1=K_1-1|\, A_2=2,A_3\neq 2)|\, (A_1=1|\, A_2=2,A_3\neq 2))}(\delta) = \frac{p^{\mathrm{k-1}}_{(N_1=K_1-1\,|A_2A_2=2,A_3\neq 2)}(\delta)}{p^{\mathrm{k-1}}_{(A_1=1|A_2=2\,,A_3\neq 2)}(\delta)}$$

$$\alpha^{e^{k}}_{3} = p^{\mathrm{k-1}}_{((N_1=1|\, A_2=2\,,A_3=2)|(A_1=1|\, A_2=2\,,A_3=2))}(\delta) = \frac{p^{\mathrm{k-1}}_{(N_1=1\,|A_2=2\,,A_3=2)}(\delta)}{p^{\mathrm{k-1}}_{(A_1=1|A_2=2\,,A_3=2)}(\delta)}$$

$$\alpha^{f^{k}}_{3} = p^{\mathrm{k-1}}_{((N_1=K_1-1|\, A_2=2\,,A_3=2)|(A_1=1|\, A_2=2\,,A_3=2))}(\delta) = \frac{p^{\mathrm{k-1}}_{(N_1=K_1-1\,|A_2=2\,,A_3=2)}(\delta)}{p^{\mathrm{k-1}}_{(A_1=1|A_2=2\,,A_3=2)}(\delta)}$$

$$\alpha^{e^{k}}_{4} = p^{\mathrm{k-1}}_{((N_2=1|A_3\neq 2)|(A_2=1|A_3\neq 2))}(\delta) = \frac{p^{\mathrm{k-1}}_{(N_2=1\,|A_3\neq 2)}(\delta)}{p^{\mathrm{k-1}}_{(A_2=1|A_3\neq 2)}(\delta)}$$

$$
\begin{gathered}
\alpha_4^{f^k} = p^{\mathrm{k-1}}_{((N_2=K_2-1|A_3\neq 2)|(A_2=1|A_3\neq 2)}(\delta) = \frac{p^{\mathrm{k-1}}_{(N_2=K_2-1\ |A_3\neq 2)}(\delta)}{p^{\mathrm{k-1}}_{(A_2=1|\ A_3\neq 2)}(\delta)} \\
\alpha_5^{e^k} = p^{\mathrm{k-1}}_{((N_2=1|A_3=2)\ |(A_2=1|A_3=2))}(\delta) = \frac{p^{\mathrm{k-1}}_{(N_2=1\ |A_3=2)}(\delta)}{p^{\mathrm{k-1}}_{(A_2=1|A_3=2)}(\delta)} \\
\alpha_5^{f^k} = p^{\mathrm{k-1}}_{((N_2=K_2-1|A_3=2)\ |(A_2=1|A_3=2))}(\delta) = \frac{p^{\mathrm{k-1}}_{(N_2=K_2-1\ |A_3=2)}(\delta)}{p^{\mathrm{k-1}}_{(A_2=1|A_3=2)}(\delta)} \\
\alpha_6^{e^k} = p^{\mathrm{k-1}}_{(N_3=1\ |A_3=1)}(\delta) = \frac{p^{\mathrm{k-1}}_{(N_3=1)}(\delta)}{p^{\mathrm{k-1}}_{(A_3=1)}(\delta)} \\
\alpha_6^{f^k} = p^{\mathrm{k-1}}_{(N_3=K_3-1\ |A_3=1)}(\delta) = \frac{p^{\mathrm{k-1}}_{(N_3=K_3-1)}(\delta)}{p^{\mathrm{k-1}}_{(A_3=1)}(\delta)}
\end{gathered}
\tag{18}
$$

For each of the 6 blocking scenarios in Table 2-1, at time step *k,* we define 2 aggregate transition rate probabilities, the full aggregate transition rate probability, denoted $\alpha_j^{f^k}$, and the empty aggregate transition rate probability, denoted $\alpha_j^{e^k}$. $\alpha_j^{e^k}$ represents the ratio of the probability of being in disaggregate state 1 given blocking scenario *j* to the probability of being in aggregate state 1 given blocking scenario *j* at time interval *k-1.* While $\alpha_j^{f^k}$ represents the ratio of the probability of being in disaggregate state K_j-1 given blocking scenario *j* to the probability of being in aggregate state 1 given blocking scenario *j* at time interval *k-1.*

Calculating the full and empty aggregate transition rate probabilities for all the blocking scenarios, defined in Equation (18), is somewhat of a challenge given that the disaggregate probabilities in the numerators are unknown. To approximate the disaggregate probabilities, we follow the same approach as in the one queue model. That is by assuming the disaggregate probabilities of the blocking scenarios follow the functional form given by Morse (1958) in Equation (12). We solve 6 different nonlinear systems for all blocking scenarios. We solve the nonlinear systems in Equations (19) through (24) to obtain six different pairs of λ_j^{k-1} and μ_j^{k-1}, where j represents the blocking scenario index.

To approximate the disaggregate probabilities for blocking scenario 1: $p^{\mathrm{k}-1}_{(N_1=1|A_2\neq 2)}(\delta)$, $p^{\mathrm{k}-1}_{(N_1=K_1-1|A_2\neq 2)}(\delta)$, during time interval $k-1$, we solve the following nonlinear system for λ_1^{k-1} and μ_1^{k-1}:

$$p^{\mathrm{k}-1}_{(A_1=0|A_2\neq 2)}(\delta) - \left(\sum_{m=0}^{K_1} p^{k-2}_{(N_1=m|A_2\neq 2)}(\delta)\ d^{m,k-1}_{N=0}(\delta) \right) = 0,$$

$$p^{k-1}_{(A_1=2|A_2\neq 2)}(\delta) - \left(\sum_{m=0}^{K_1} p^{k-2}_{(N_1=m|A_2\neq 2)}(\delta)\ \ d^{m,k-1}_{N=K_1}(\delta) \right) = 0.$$

(19)

To approximate the disaggregate probabilities for blocking scenario 2: $p^{k-1}_{(N_1=1\ |A_2=2,A_3\neq 2)}(\delta)$, $p^{k-1}_{(N_1=K_1-1\ |A_2=2,A_3\neq 2)}(\delta)$ during time interval $k-1$, we solve the following nonlinear system for λ_2^{k-1} and μ_2^{k-1}:

$$p^{k-1}_{(A_1=0\ |A_2=2,A_3\neq 2)}(\delta) - \left(\sum_{m=0}^{K_1} p^{k-2}_{(N_1=m|A_2=2,A_3\neq 2)}\ (\delta)\ d^{m,k-1}_{N=0}(\delta) \right) = 0,$$

$$p^{k-1}_{(A_1=2\ |A_2=2,A_3\neq 2)}(\delta) - \left(\sum_{m=0}^{K_1} p^{k-2}_{(N_1=m|A_2=2,A_3\neq 2)}\ (\delta)\ d^{m,k-1}_{N=K_1}(\delta) \right) = 0.$$

(20)

To approximate the disaggregate probabilities for blocking scenario 3: $p^{k-1}_{(N_1=1\ |A_2=2,A_3=2)}(\delta)$, $p^{k-1}_{(N_1=K_1-1\ |A_2=2,A_3=2)}(\delta)$ during time interval $k-1$, we solve the following nonlinear system for λ_3^{k-1} and μ_3^{k-1}:

$$p^{k-1}_{(A_1=0\ |A_2=2,A_3=2)}(\delta) - \left(\sum_{m=0}^{K_1} p^{k-2}_{(N_1=m|A_2=2,A_3=2)}\ (\delta)\ \ d^{m,k-1}_{N=0}(\delta) \right) = 0,$$

$$p^{k-1}_{(A_1=2\ |A_2=2,A_3=2)}(\delta)-\left(\sum_{m=0}^{K_1} p^{k-2}_{(N_1=m|A_2=2,A_3\neq 2)}\ (\delta)\ d^{m,k-1}_{N=K_1}(\delta)\right)=0.$$

(21)

To approximate the disaggregate probabilities for blocking scenario 4: $p^{k-1}_{(N_2=1\ |A_3\neq 2)}(\delta)$, $p^{k-1}_{(N_2=K_2-1\ |A_3\neq 2)}(\delta)$ during time interval $k-1$, we solve the following nonlinear system for λ_4^{k-1} and μ_4^{k-1}:

$$p^{k-1}_{(A_2=0\ |A_3\neq 2)}(\delta)-\left(\sum_{m=0}^{K_2} p^{k-2}_{(N_2=m\ |A_3\neq 2)}(\delta)\ \ d^{m,k-1}_{N=0}(\delta)\right)=0,$$

$$p^{k-1}_{(A_2=2\ |A_3\neq 2)}(\delta)-\left(\sum_{m=0}^{K_2} p^{k-2}_{(N_2=m\ |A_3\neq 2)}\ (\delta)\ d^{m,k-1}_{N=K_1}(\delta)\right)=0,$$

(22)

To approximate the disaggregate probabilities for blocking scenario 5: $p^{k-1}_{(N_2=1\ |A_3=2)}(\delta)$, $p^{k-1}_{(N_2=K_2-1\ |A_3=2)}(\delta)$ during time interval $k-1$, we solve the following nonlinear system for λ_5^{k-1} and μ_5^{k-1}:

$$p^{k-1}_{(A_2=0\ |A_3=2)}(\delta)-\left(\sum_{m=0}^{K_2} p^{k-2}_{(N_2=m\ |A_3=2)}\ (\delta)\ d^{m,k-1}_{N=0}(\delta)\right)=0,$$

$$p^{k-1}_{(A_2=2\ |A_3=2)}(\delta)-\left(\sum_{m=0}^{K_2} p^{k-2}_{(N_2=m\ |A_3=2)}(\delta)\ \ d^{m,k-1}_{N=K_1}(\delta)\right)=0.$$

(23)

To approximate the disaggregate probability for blocking scenario 6: $p^{k-1}_{(N_3=1)}(\delta)$, $p^{k-1}_{(N_3=K_3-1)}(\delta)$, during time interval $k-1$, we solve the following nonlinear system for λ_6^{k-1} and μ_6^{k-1}:

$$p^{k-1}_{(A_3=0)}(\delta)-\left(\sum_{m=0}^{K_3} p^{k-2}_{(N_3=m)}(\delta)\ \ d^{m,k-1}_{N=0}(\delta)\right)=0,$$

$$p^{k-1}_{(A_3=2)}(\delta) - \left(\sum_{m=0}^{K_3} p^{k-2}_{(N_3=m)}\,(\delta)\; d^{m,k-1}_{N=K_1}(\delta)\right) = 0. \tag{24}$$

For Equations (19) through (24), we calculate $d^{m,k-1}_{N=0}(\delta), d^{m,k-1}_{N=K_1}(\delta)$ from equation (12b-2). Once λ_j^{k-1} and μ_j^{k-1} for $j\,\epsilon\{1,2,3,4,5,6\}$ are obtained from the nonlinear solver, we plug them into the discrete form of Equation (12) to get the disaggregate probabilities needed. These steps are described in more details in the algorithm description below.

The full algorithm for solving the transient joint distribution of a three-queue network can be described in the following steps:

Input:

External arrival rates to each of the three queues : $\gamma = [\gamma_1, \gamma_2, \gamma_3]$.

Service rate for each of the three queues: $\mu = [\mu_1, \mu_2, \mu_3]$.

Queue capacity for each of the three queues $K = [K_1, K_2, K_3]$.

Initial disaggregate distribution for each of the three queues: $p^0_{N_1}, p^0_{N_2}, p^0_{N_3}$.

Duration of entire time horizon: T.

Output :

Assuming that $\frac{T}{\delta}$ is an integer, the output is an approximation of the aggregate queue-length distribution of a queue at time T (in discrete time at time interval $\frac{T}{\delta}$):

$p^{\frac{T}{\delta}}_{A=(i,j,l)}(t,)$ where t can be any value between $[0, \delta]$. .

Algorithm:

δ can be initiated as any small number.

```
For time step $k = 0,1,2,\dots,\left\lceil\frac{T}{\delta}\right\rceil$
    If  $k = 0$
```

1) Calculate the initial aggregate distribution ($p_{A_i}^0$) from the initial disaggregate distribution ($p_{N_i}^0$) for $i\epsilon\{1,2,3\}$ using the following equation:

$$p_{A_i}^0 = \begin{bmatrix} p_{N_i=0}^0 \\ 1 - p_{N_i=0}^0 - p_{N_i=K}^0 \\ p_{N_i=K}^0 \end{bmatrix},$$

2) Calculate the initial joint queue-length distribution from the following equation:

$$p_{A=(i,j,l)}^0 = p_{A_1=i}^0 \, p_{A_2=j}^0 p_{A_3=l}^0$$

3) Calculate the initial aggregate transition rates for each of the blocking scenarios $\alpha_i^{e1}, \alpha_i^{f1}$ for $i \epsilon \{1,2,3,4,5,6\}$ from the initial joint $p_{A=(i,j,l)}^0$ and the initial disaggregate distributions $p_{N_1}^0, p_{N_2}^0, p_{N_3}^0$:

$$\alpha_1^{e1} = \frac{p_{(N_1=1 \,|A_2\neq 2)}^0}{p_{(A_1=1|A_2\neq 2)}^0} , \alpha_1^{f1} = \frac{p_{(N_1=K_1-1 \,|A_2\neq 2)}^0}{p_{(A_1=1|A_2\neq 2)}^0}$$

$$\alpha_2^{e1} = \frac{p_{(N_1=1 \,|A_2=2,A_3\neq 2)}^0}{p_{(A_1=1|\, A_2=2,A_3\neq 2)}^0} , \alpha_2^{f1} = \frac{p_{(N_1=K_1-1 \,|A_2 A_2=2,A_3\neq 2)}^0}{p_{(A_1=1|A_2=2 \,,A_3\neq 2)}^0}$$

$$\alpha_3^{e1} = \frac{p_{(N_1=1 \,|A_2=2 \,,A_3=2)}^0}{p_{(A_1=1|A_2=2 \,,A_3=2)}^0} , \alpha_3^{f1} = \frac{p_{(N_1=K_1-1 \,|A_2=2 \,,A_3=2)}^0}{p_{(A_1=1|A_2=2 \,,A_3=2)}^0}$$

$$\alpha_4^{e1} = \frac{p_{(N_2=1 \,|A_3\neq 2)}^0}{p_{(A_2=1|A_3\neq 2)}^0}, \alpha_4^{f1} = \frac{p_{(N_2=K_2-1 \,|A_3\neq 2)}^0}{p_{(A_2=1|\, A_3\neq 2)}^{\mathrm{k}-1}}$$

$$\alpha_5^{e1} = \frac{p_{(N_2=1 \,|A_3=2)}^0}{p_{(A_2=1|A_3=2)}^{\mathrm{k}-1}} , \alpha_5^{f1} = \frac{p_{(N_2=K_2-1 \,|A_3=2)}^0}{p_{(A_2=1|A_3=2)}^0}$$

$$\alpha_6^{e1} = \frac{p_{(N_3=1)}^0}{p_{(A_3=1)}^0} , \alpha_6^{f1} = \frac{p_{(N_3=K_3-1)}^0}{p_{(A_3=1)}^0}$$

Else

1) The aggregate joint queue-length distribution for time step k of continuous time t is defined as:

$$p^k_{A=(i,j,l)}(t) = p^{k-1}_{A=(i,j,l)}(\delta)e^{t\,Q^k_{AJ}}, \quad \forall t \in [0,\delta]\,,$$

where $Q^k_{AJ} = f(\gamma,\mu,\alpha^k,B)$ is a 27x27 sparse matrix with nonzero elements described in appendix A. The parameters for the matrix are:
$\gamma = [\gamma_1,\gamma_2,\gamma_3]$, $\mu = [\mu_1,\mu_2,\mu_3]$ given initially as input ,
$B = [B_1,B_2,B_3,B_4]$ given in Table 2-1, and
$\alpha^k = [\alpha_1^{ek},\alpha_1^{f^k},\alpha_2^{ek},\alpha_2^{f^k},\alpha_3^{ek},\ \alpha_3^{f^k},\ \alpha_4^{ek},\ \alpha_4^{f^k},\ \alpha_5^{ek},\ \alpha_5^{f^k},\ \alpha_6^{ek},\ \alpha_6^{f^k}]$
approximated in the previous time step.

2) Solve six nonlinear system of equations for the six blocking scenarios to obtain λ^k_j,μ^k_j where j is the blocking scenario index:

Nonlinear system 1: Solve to obtain λ^k_1,μ^k_1

$$p^{\mathrm{k}}_{(A_1=0|A_2\neq2)}(\delta) - \left(\sum_{m=0}^{K_1} p^{k-1}_{(N_1=m|A_2\neq2)}(\delta)\ d^{m,k}_{N=0}(\delta)\right) = 0$$

$$p^{k}_{(A_1=2|A_2\neq2)}(\delta) - \left(\sum_{m=0}^{K_1} p^{k-1}_{(N_1=m|A_2\neq2)}(\delta)\ \ d^{m,k}_{N=K_1}(\delta)\right) = 0$$

Nonlinear system 2: Solve to obtain λ^k_2,μ^k_2

$$p^{k}_{(A_1=0\ |A_2=2,A_3\neq2)}(\delta) - \left(\sum_{m=0}^{K_1} p^{k-1}_{(N_1=m|A_2=2,A_3\neq2)}(\delta)\ d^{m,k}_{N=0}(\delta)\right) = 0$$

$$p^{k}_{(A_1=2\ |A_2=2,A_3\neq2)}(\delta) - \left(\sum_{m=0}^{K_1} p^{k-1}_{(N_1=m|A_2=2,A_3\neq2)}(\delta)d^{m,k}_{N=K_1}(\delta)\right) = 0$$

Nonlinear system 3: Solve to obtain λ^k_3,μ^k_3

$$p^{k}_{(A_1=0\ |A_2=2,A_3=2)}(\delta) - \left(\sum_{m=0}^{K_1} p^{k-1}_{(N_1=m|A_2=2,A_3=2)}(\delta)\ d^{m,k}_{N=0}(\delta)\right) = 0$$

$$p^{k}_{(A_1=2\ |A_2=2,A_3=2)}(\delta) - \left(\sum_{m=0}^{K_1} p^{k-1}_{(N_1=m|A_2=2,A_3\neq 2)}\ (\delta)\ d^{m,k}_{N=K_1}(\delta) \right) = 0$$

Nonlinear system 4: Solve to obtain λ_4^k, μ_4^k

$$p^{k}_{(A_2=0\ |A_3\neq 2)}(\delta) - \left(\sum_{m=0}^{K_2} p^{k-1}_{(N_2=m\ |A_3\neq 2)}(\delta)\ \ d^{m,k}_{N=0}(\delta) \right) = 0$$

$$p^{k}_{(A_2=2\ |A_3\neq 2)}(\delta) - \left(\sum_{m=0}^{K_2} p^{k-1}_{(N_2=m\ |A_3\neq 2)}\ (\delta)\ d^{m,k}_{N=K_1}(\delta) \right) = 0$$

Nonlinear system 5: Solve to obtain λ_5^k, μ_5^k

$$p^{k}_{(A_2=0\ |A_3=2)}(\delta) - \left(\sum_{m=0}^{K_2} p^{k-1}_{(N_2=m\ |A_3=2)}\ (\delta)\ d^{m,k}_{N=0}(\delta) \right) = 0$$

$$p^{k}_{(A_2=2\ |A_3=2)}(\delta) - \left(\sum_{m=0}^{K_2} p^{k-1}_{(N_2=m\ |A_3=2)}(\delta)\ \ d^{m,k}_{N=K_1}(\delta) \right) = 0$$

Nonlinear system 6: Solve to obtain λ_6^k, μ_6^k

$$p^{k}_{(A_3=0)}(\delta) - \left(\sum_{m=0}^{K_3} p^{k-1}_{(N_3=m)}(\delta)\ \ d^{m,k}_{N=0}(\delta) \right) = 0$$

$$p^{k}_{(A_3=2)}(\delta) - \left(\sum_{m=0}^{K_3} p^{k-1}_{(N_3=m)}\ (\delta)\ d^{m,k}_{N=K_1}(\delta) \right) = 0$$

3) Plug each pair λ_j^k, μ_j^k and the disaggregate distribution for each blocking scenario from time step $k-1$ as the initial distribution into the discrete form of Equation (12) to get the disaggregate probabilities of being in disaggregate states 1, $K-1$.

   Plug λ_1^k, μ_1^k and the disaggregate distribution for this blocking scenario from time step $k-1$ as the initial distribution, into Equation (12) to get:

$$p^{\mathrm{k}}_{(N_1=1|A_2\neq2)}(\delta) = \sum_{m=0}^{K_1} p^{k-1}_{(N_1=m|A_2\neq2)}(\delta) d^{m,k}_{N=1}(\delta),$$

$$p^{\mathrm{k}}_{(N_1=K_1-1|A_2\neq2)}(\delta) = \sum_{m=0}^{K_1} p^{k-1}_{(N_1=m|A_2\neq2)}(\delta) d^{m,k}_{N=K_1-1}(\delta).$$

Plug λ_2^k, μ_2^k into equation (12) to get

$$p^{k}_{(N_1=1\ |A_2=2,A_3\neq2)}(\delta) = \sum_{m=0}^{K_1} p^{k-1}_{(N_1=m|A_2=2,A_3\neq2)}(\delta)\ d^{m,k}_{N=1}(\delta),$$

$$p^{k}_{(N_1=K_1-1\ |A_2=2,A_3\neq2)}(\delta) = \sum_{m=0}^{K_1} p^{k-1}_{(N_1=m|A_2=2,A_3\neq2)}(\delta)\ d^{m,k}_{N=K_1-1}(\delta).$$

Plug λ_3^k, μ_3^k into equation (12) to get

$$p^{\mathrm{k}}_{(N_1=1|A_2=2,A_3=2)}(\delta) = \sum_{m=0}^{K_1} p^{k-1}_{(N_1=m|A_2=2,A_3=2)}(\delta) d^{m,k}_{N=1}(\delta),$$

$$p^{\mathrm{k}}_{(N_1=K_1-1|A_2=2,A_3=2)}(\delta) = \sum_{m=0}^{K_1} p^{k-1}_{(N_1=m|A_2=2,A_3=2)}(\delta) d^{m,k}_{N=K_1-1}(\delta).$$

Plug λ_4^k, μ_4^k into equation (12) to get

$$p^{k}_{(N_2=1|A_3\neq2)}(\delta) = \sum_{m=0}^{K_2} p^{k-1}_{(N_2=m|A_3\neq2)}(\delta) d^{m,k}_{N=1}(\delta),$$

$$p^{\mathrm{k}}_{(N_2=K_2-1|A_3\neq2)}(\delta) = \sum_{m=0}^{K_2} p^{k-1}_{(N_2=m|A_3\neq2)}(\delta) d^{m,k}_{N=K_2-1}(\delta).$$

Plug λ_5^k, μ_5^k into equation (12) to get

$$p^{\mathrm{k}}_{(N_2=1|A_3=2)}(\delta) = \sum_{m=0}^{K_2} p^{k-1}_{(N_2=m|A_3=2)}(\delta) d^{m,k}_{N=1}(\delta),$$

$$p^{\mathrm{k}}_{(N_2=K_2-1|A_3=2)}(\delta) = \sum_{m=0}^{K_2} p^{k-1}_{(N_2=m|A_3=2)}(\delta) d^{m,k}_{N=K_2-1}(\delta).$$

Plug λ_6^k, μ_6^k into equation (12) to get

$$p^{k}_{(N_3=1)}(\delta) = \sum_{m=0}^{K_3} p^{k-1}_{(N_3=m)}(\delta) d^{m,k}_{N=1}(\delta),$$

$$p^{k}_{(N_3=K_3-1)}(\delta) = \sum_{m=0}^{K_3} p^{k-1}_{(N_3=m)}(\delta) d^{m,k}_{N=K_3-1}(\delta).$$

4) Calculate $\alpha_j^{e^{k+1}}, \alpha_j^{f^{k+1}}$, where j is the blocking scenario for time step $k+1$ from the following equations:

$$\alpha_1^{e^{k+1}} = \frac{p^{\mathrm{k}}_{(N_1=1\ |A_2\neq 2)}(\delta)}{p^{\mathrm{k}}_{(A_1=1|A_2\neq 2)}(\delta)}\ , \alpha_1^{f^{k+1}} = \frac{p^{\mathrm{k}}_{(N_1=K_1-1\ |A_2\neq 2)}(\delta)}{p^{\mathrm{k}}_{(A_1=1|A_2\neq 2)}(\delta)}$$

$$\alpha_2^{e^{k+1}} = \frac{p^{\mathrm{k}}_{(N_1=1\ |A_2=2,A_3\neq 2)}(\delta)}{p^{\mathrm{k}}_{(A_1=1|\ A_2=2,A_3\neq 2)}(\delta)}\ , \alpha_2^{f^{k+1}} = \frac{p^{\mathrm{k}}_{(N_1=K_1-1\ |A_2=2,A_3\neq 2)}(\delta)}{p^{\mathrm{k}}_{(A_1=1|A_2=2\ ,A_3\neq 2)}(\delta)}$$

$$\alpha_3^{e^{k+1}} = \frac{p^{\mathrm{k}}_{(N_1=1\ |A_2=2\ ,A_3=2)}(\delta)}{p^{\mathrm{k}}_{(A_1=1|A_2=2\ ,A_3=2)}(\delta)}\ , \alpha_3^{f^{k+1}} = \frac{p^{\mathrm{k}}_{(N_1=K_1-1\ |A_2=2\ ,A_3=2)}(\delta)}{p^{\mathrm{k}}_{(A_1=1|A_2=2\ ,A_3=2)}(\delta)}$$

$$\alpha_4^{e^{k+1}} = \frac{p^{\mathrm{k}}_{(N_2=1\ |A_3\neq 2)}(\delta)}{p^{\mathrm{k}}_{(A_2=1|A_3\neq 2)}(\delta)}\ , \alpha_4^{f^{k+1}} = \frac{p^{\mathrm{k}}_{(N_2=K_2-1\ |A_3\neq 2)}(\delta)}{p^{\mathrm{k}}_{(A_2=1|\ A_3\neq 2)}(\delta)}$$

$$\alpha_5^{e^{k+1}} = \frac{p^{\mathrm{k}}_{(N_2=1\ |A_3=2)}(\delta)}{p^{\mathrm{k}}_{(A_2=1|A_3=2)}(\delta)}\ , \alpha_5^{f^{k+1}} = \frac{p^{\mathrm{k}}_{(N_2=K_2-1\ |A_3=2)}(\delta)}{p^{\mathrm{k}}_{(A_2=1|A_3=2)}(\delta)}$$

$$\alpha_6^{e^{k+1}} = \frac{p^{\mathrm{k}}_{(N_3=1)}(\delta)}{p^{\mathrm{k}}_{(A_3=1)}(\delta)}\ , \alpha_6^{f^{k+1}} = \frac{p^{\mathrm{k}}_{(N_3=K_3-1)}(\delta)}{p^{\mathrm{k}}_{(A_3=1)}(\delta)}$$

```
        End
End
```

2.2.3 Aggregate transient model for an M-queue tandem network

We generalize the method of computing the transient queue-length distributions for M queues in tandem by decomposing the network into overlapping 3-queue sub-networks, illustrated in Figure 2-2. The method that we will apply for each of the sub-networks is based on the one developed in the previous section. This approach not only maintains the same level of linear computational complexity that we mentioned in the previous section, but also allows us to validate the accuracy of marginal transient distributions for individual queues. The total number of sub-networks that we need to evaluate is M-2.

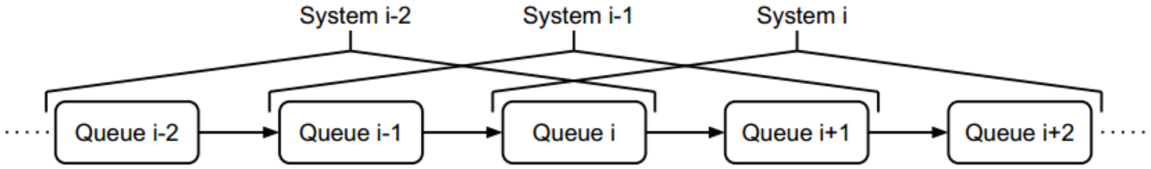


Figure 2-2: Decomposing the network to overlapping sub-networks of three tandem queues (Osorio and Wang, 2012)

To present the model, we introduce the following notation:

q_i	queue *i*;
N_i	disaggregate state of queue *i*;
A_i	aggregate state of queue *i*;
γ_i	external arrival rate to queue *i*;
μ_i	exogenous service rate of queue *i*;
K_i	capacity of queue *i*;
K_j	capacity of the queue corresponding to blocking scenario *j*;
λ_i^k	total arrival rate to queue *i* during time interval *k*;

$\widehat{\mu_i}^k$ effective service rate of queue *i* during time interval *k*;

$\widetilde{\mu_i^k}$ unblocking rate of queue *i* during time interval *k;*

${p^{b}}^{k}_{i}$ blocking probability of queue *i* during time interval *k;*

$\lambda_{s,j}^k$ approximated queue arrival rate for sub-network *s* and blocking scenario *j* during time interval *k*;

$\mu_{s,j}^k$ approximated queue arrival rate for sub-network *s* and blocking scenario *j* during time interval *k*;

$\rho_{s,j}^k$ approximated queue traffic intensity for sub-network *s* and blocking scenario *j* during time interval *k*;

${\alpha_{s,j}^{e}}^{k}$ empty aggregate transition rate probability for sub-network *s* and blocking scenario *i* during time interval k;

${\alpha_{s,j}^{f}}^{k}$ full aggregate transition rate probability for sub-network *s* and blocking scenario *i* during time interval k;

$Q_{AJ_s}^k$ aggregate joint transition rate matrix for sub-network *s* during time interval k;

$p_{N_i}^0$ initial disaggregate queue-length distribution for queue *i*;

$p_{A_i}^0$ initial aggregate queue-length distribution for queue *i*;

$p_{s,A_i=a}^k(t)$ the marginal probability that queue *i* in sub-network *s* is in aggregate state *a* at continuous time *t* within time interval k;

$p_{s,N_i=n}^k(t)$ the marginal probability that queue *i* in sub-network *s* is in disaggregate state *n* at continuous time *t* within time interval k.

$p_{s,A=(i,j,l)}^0$ initial aggregate joint queue-length distribution for sub-network *s*;

$p_{s,N=(i,j,l)}^0$ initial disaggregate joint queue-length distribution for sub-network *s*;

$p_{s,A=(i,j,l)}^k(t)$ aggregate joint queue-length distribution of sub-network *s* at continuous time t within time interval k;

$p_{s,N=(i,j,l)}^k(t)$ disaggregate joint queue-length distribution of sub-network *s* at continuous time t within time interval k;

δ time step length;

T duration of entire time horizon of which the joint queue-length distribution of the M-queue network is evaluated;

t continuous time within the $[0, \delta]$ interval.

For any sub-network *i*, with queue indices $(i, i+1\,, i+2)$, to calculate an accurate joint queue-length distribution, we need to understand the dependencies between adjacent queues to the sub-networks and the effects of both upstream and downstream queues. The adjacent upstream queue *i-1* gives us information on the arrival rate of queue *i,* and the adjacent downstream queue *i+3* gives us information on the service rate of queue *i+2*. Hence, for each sub-network *i*, the total arrival rate to the first queue λ_i^k, and the effective service rate of the third queue $\widehat{\mu_{i+2}}^k$, during time interval *k*, is calculated by using information from adjacent queues. The total arrival rate to the most upstream queue in system *i,* queue *i*, is obtained by solving the flow conservation equation derived by Osorio and Bierlaire (2009a) and given by:

$$\lambda_i^k = \gamma_i + \frac{\lambda_{i-1}^k(1-p^k_{A_{i-1}=2}(0))}{(1-p^k_{A_i=2}(0))}. \tag{25}$$

The effective service rate, $\widehat{\mu_{i+2}}^k$, for the most downstream queue in system *i,* queue *i+2,* accounts for service and for potential blocking from downstream queues. It is also derived by Osorio and Bierlaire (2009a), and given by:

$$\widehat{\mu_i}^k = \left(\frac{1}{\mu_i} + {p^b}_i^k \frac{1}{\widetilde{\mu_i^k}}\right)^{-1}, \tag{26}$$

where μ_i is the exogenous service rate, ${p^b}_i^k$is the blocking probability during time interval k, and $\widetilde{\mu_i^k}$ is the unblocking rate during time interval *k* of q_i. The approximation for $\widetilde{\mu_i^k}$ for a single queue is derived by Osorio and Bierlaire (2009b), and is given by:

$$\frac{1}{\widetilde{\mu_i^k}} = \frac{\lambda_{i+1}^k\left(1-p^k_{A_{i+1}=2}(0)\right)}{\lambda_i^k\left(1-p^k_{A_i=2}(0)\right)}\,\frac{1}{\hat{\mu}_{i+1}}.$$

(27)

Additionally, ${p^{b}}_{i}^{k}$ is approximated by:

$${p^{b}}_{i}^{k} = p^{k}_{(A_{i+1}=2)}(0)\ \frac{\mu_i}{\mu_i + \mu_{i+1}}.$$

(28)

We substitute Equations (27) and (28) in (26) and get

$$\widehat{\mu_i}^{k} = \left(\frac{1}{\mu_i} + p^{k}_{(A_{i+1}=2)}(0)\ \frac{\mu_i}{\mu_i + \mu_{i+1}}\ \frac{\lambda^{k}_{i+1}\left(1 - p^{k}_{A_{i+1}=2}(0)\right)}{\lambda^{k}_{i}\left(1 - p^{k}_{A_i=2}(0)\right)}\ \frac{1}{\hat{\mu}_{i+1}}\right)^{-1}.$$

(29)

The other important aspect to consider for this method is the consistency of marginal queue-length distributions of same queues in different sub-networks. Our method does not ensure consistency among same queue marginal distributions in different sub-network. However, we ensure consistency among the aggregate transition rate probabilities for the same queues in different sub-networks through system of Equations (30).

$$\alpha^{e^k}_{i,4} = \alpha^{e^k}_{i+1,1}$$

(30a)

$$\alpha^{f^k}_{i,4} = \alpha^{f^k}_{i+1,1}$$

(30b)

$$\alpha^{e^k}_{i,5} = p^{k-1}_{i+1,(A_{i+3}\neq 2)}(\delta)\ \alpha^{e^k}_{i+1,2} + p^{k-1}_{i+1,(A_{i+3}=2)}(\delta)\ \alpha^{e^k}_{i+1,3}$$

(30c)

$$\alpha^{f^k}_{i,5} = p^{k-1}_{i+1,(A_{i+3}\neq 2)}(\delta)\ \alpha^{f^k}_{i+1,2} + p^{k-1}_{i+1,(A_{i+3}=2)}(\delta)\ \alpha^{f^k}_{i+1,3}$$

(30d)

$$\begin{aligned}\alpha_{i,6}^{e^k} &= p_{i+1,(A_{i+3}\neq 2)}^{k-1}(\delta)\, \alpha_{i+1,4}^{e^k} + p_{i+1,(A_{i+3}=2)}^{k-1}(\delta)\, \alpha_{i+1,5}^{e^k} = \\ & p_{i+1,(A_{i+3}\neq 2)}^{k-1}(\delta)\, \alpha_{i+2,1}^{e^k} + p_{i+1,(A_{i+3}=2)}^{k-1}(\delta)\, (p_{i+2,(A_{i+4}\neq 2)}^{k-1}(\delta)\, \alpha_{i+2,2}^{e^k} \\ & + p_{i+2,(A_{i+4}=2)}^{k-1}(\delta)\, \alpha_{i+2,3}^{e^k})\end{aligned} \tag{30e}$$

$$\begin{aligned}\alpha_{i,6}^{f^k} &= p_{i+1,(A_{i+3}\neq 2)}^{k-1}(\delta)\, \alpha_{i+1,4}^{f^k} + p_{i+1,(A_{i+3}=2)}^{k-1}(\delta)\, \alpha_{i+1,5}^{f^k} = \\ & p_{i+1,(A_{i+3}\neq 2)}^{k-1}(\delta)\, \alpha_{i+2,1}^{f^k} + p_{i+1,(A_{i+3}=2)}^{k-1}(\delta)\, (p_{i+2,(A_{i+4}\neq 2)}^{k-1}(\delta)\, \alpha_{i+2,2}^{f^k} \\ & + p_{i+2,(A_{i+4}=2)}^{k-1}(\delta)\, \alpha_{i+2,3}^{f^k})\end{aligned} \tag{30f}$$

If we have two overlapping sub-networks, *i* and *i+1,* then q_{i+1} is the second queue of sub-network *i* and the first queue of sub-network *i+1*. In sub-network *i*, q_{i+1} has two blocking scenarios: q_{i+1} is not blocked and q_{i+1} is blocked by q_{i+2}. The aggregate transition rate probabilities for these scenarios are $\alpha_{i,4}^{e^k}, \alpha_{i,4}^{f^k}, \alpha_{i,5}^{e^k}, \alpha_{i,5}^{f^k}$. The same queue q_{i+1} of the sub-network *i+1* has instead three blocking scenarios: q_{i+1} not blocked, q_{i+1} blocked by q_{i+2} , and q_{i+1} blocked by q_{i+3} . The aggregate transition rate probabilities for these scenarios are $\alpha_{i+1,1}^{e^k}, \alpha_{i+1,1}^{f^k}, \alpha_{i+1,2}^{e^k}, \alpha_{i+1,2}^{f^k}, \alpha_{i+1,3}^{e^k}, \alpha_{i+1,3}^{f^k}$. Equation (30a) shows that $\alpha_{i,4}^{e^k}$ and $\alpha_{i+1,1}^{e^k}$ are equal , and equation (30b) shows that $\alpha_{i,4}^{f^k}$ and $\alpha_{i+1,1}^{f^k}$ are equal, since they are the probabilities for the scenario that q_{i+1} is not blocked by q_{i+2}.

Additionally, $\alpha_{i,5}^{e^k}, \alpha_{i,5}^{f^k}$ are the probabilities for the scenario that q_{i+1} is blocked by q_{i+2} whereas $\alpha_{i+1,2}^{e^k}, \alpha_{i+1,2}^{f^k}, \alpha_{i+1,3}^{e^k}, \alpha_{i+1,3}^{f^k}$ are the probabilities for the scenario that q_{i+1} is blocked by q_{i+2} but conditioned upon information on q_{i+3}. Statistically, $\alpha_{i,5}^{e}$ is defined as a weighted average of $\alpha_{i+1,2}^{e^k}$, $\alpha_{i+1,3}^{e^k}$, with weights $p_{i+1,(A_{i+3}\neq 2)}^{k-1}(\delta), p_{i+1,(A_{i+3}=2)}^{k-1}(\delta)$ respectively, as defined in equation (30c). Similarly, $\alpha_{i,5}^{f^k}$ is defined as a weighted average of $\alpha_{i+1,2}^{f^k}$, $\alpha_{i+1,3}^{f^k}$, with weights $p_{i+1,(A_{i+3}\neq 2)}^{k-1}(\delta), p_{i+1,(A_{i+3}=2)}^{k-1}(\delta)$ respectively, as defined in equation (30d).

If we look at sub-networks *i,i+1,* and *i+2,* then q_{i+2} is the third queue of sub-network *i*, the second queue of sub-network *i+1,* and the first queue of sub-network *i+2*. In sub-network i, q_{i+2} has one blocking scenarios: q_{i+2} is not blocked. The aggregate transition rate probabilities for these scenarios are $\alpha_{i,6}^{e^k}, \alpha_{i,6}^{f^k}$. The same queue q_{i+2} of sub-network *i+1* has instead two blocking scenarios: q_{i+2} not blocked and q_{i+2} blocked by q_{i+3} . The aggregate transition rate probabilities for these scenarios are $\alpha_{i+1,4}^{e^k}, \alpha_{i+1,4}^{f^k}, \alpha_{i+1,5}^{e^k}, \alpha_{i+1,5}^{f^k}$. Additionally, q_{i+2} of sub-network *i+2* has instead three blocking scenarios: q_{i+2} not blocked, q_{i+2} blocked by q_{i+3}, and q_{i+2} blocked by q_{i+4} . The aggregate transition rate probabilities for these scenarios are $\alpha_{i+2,1}^{e^k}, \alpha_{i+2,1}^{f^k}, \alpha_{i+2,2}^{e^k}, \alpha_{i+2,2}^{f^k}, \alpha_{i+2,3}^{e^k}, \alpha_{i+2,3}^{f^k}$. Equation (30e) shows that statistically, $\alpha_{i,6}^{e^k}$ is the weighted sum of $\alpha_{i+1,4}^{e^k}$and $\alpha_{i+1,5}^{e^k}$ with weights $p_{i+1,(A_{i+3}\neq 2)}^{k-1}(\delta)$ and $p_{i+1,(A_{i+3}\neq 2)}^{k-1}(\delta)$. This can be explained using the same logic as in the previous paragraph. Equation (30f) defines the same relations as in (30e) but for the full aggregate transition probability instead of the empty aggregate transition probability.

The full algorithm for solving the transient joint distribution of an M-queue tandem network can be described in the following steps:

Input:

External arrival rates to each of the M queues : $\gamma = [\gamma_1, \gamma_2, \dots, \gamma_M]$

Service rate for each of the M queues: $\mu = [\mu_1, \mu_2, .., \mu_M]$

Queue capacity for each of the M queues $K = [K_1, K_2, .., K_M]$

Initial disaggregate distribution for each of the M queues: $p_{N_1}^0, p_{N_2}^0, .., p_{N_M}^0$

Duration of entire time horizon of which the joint queue-length distribution of the M-queue network is evaluate: T

Output :

Assuming that $\frac{T}{\delta}$ is an integer, the output are multiple 3-queue joint queue-length distribution at time *T* (in discrete time at time interval $\frac{T}{\delta}$)*:* $p_{s,A=(i,j,l)}^{\frac{T}{\delta}}(t)$*,* for each of the M-3 overlapping sub-networks in the M-queue network, where t can be any value between $[0, \delta]$.

Algorithm:

δ can be initiated as any small number.

For time step $k = 0,1,2, \dots, \frac{T}{\delta}$

If $k = 0$

1) Calculate the initial marginal aggregate distribution ($p_{A_i}^0$) from the initial marginal disaggregate distribution ($p_{N_i}^0$) for $i \epsilon \{1,2,3, \dots, M\}$, using the following equation:

$$p_{A_i}^0 = \begin{bmatrix} p_{N_i=0}^0 \\ 1 - p_{N_i=0}^0 - p_{N_i=K}^0 \\ p_{N_i=K}^0 \end{bmatrix},$$

2) Calculate the initial joint queue-length distribution for each sub-network $s \in \{1,2,3, \dots, M-2\}$, from the following equation:

$$p_{s,A=(i,j,l)}^0 = p_{A_s=i}^0 \, p_{A_{(s+1)}=j}^0 p_{A_{(s+2)}=l}^0$$

3) Calculate the initial aggregate transition rates for each subs-network $s \in \{1,2,3, \dots, M-2\}$, and blocking scenario $j \epsilon \{1,2,3,4,5,6\}$, $\alpha_{s,j}^{e^1}, \alpha_{s,j}^{f^1}$, from the initial joint $p_{s,A=(i,j,l)}^0$ and the initial disaggregate distributions $p_{N_s}^0, p_{N_{(s+1)}}^0, p_{N_{(s+2)}}^0$:

.

$$\alpha_{s,1}^{e^1} = \frac{p_{(N_s=1 \,|A_{(s+1)} \neq 2)}^0}{p_{(A_s=1|A_{(s+1)} \neq 2)}^0}, \alpha_{s,1}^{f^1} = \frac{p_{(N_s=K_s-1 \,|A_{(s+1)} \neq 2)}^0}{p_{(A_s=1|A_{(s+1)} \neq 2)}^0}$$

$$\alpha_{s,2}^{e^1} = \frac{p_{(N_s=1 \,|A_{(s+1)}=2, A_{(s+2)} \neq 2)}^0}{p_{(A_s=1| \, A_{(s+1)}=2, A_{(s+2)} \neq 2)}^0}, \alpha_{s,2}^{f^1} = \frac{p_{(N_s=K_s-1 \,|A_{(s+1)}=2, A_{(s+2)} \neq 2)}^0}{p_{(A_s=1|A_{(s+1)}=2, A_{(s+2)} \neq 2)}^0}$$

$$\alpha_{s,3}^{e^1} = \frac{p_{(N_1=1 \,|A_{(s+1)}=2, A_{(s+2)}=2)}^0}{p_{(A_1=1|A_{(s+1)}=2, A_{(s+2)}=2)}^0}, \alpha_{s,3}^{f^1} = \frac{p_{(N_s=K_s-1 \,|A_{(s+1)}=2, A_{(s+2)}=2)}^0}{p_{(A_s=1|A_{(s+1)}=2, A_{(s+2)}=2)}^0}$$

$$\alpha_{s,4}^{e^1} = \frac{p^0_{(N_{(s+1)}=1\,|A_{(s+2)}\neq 2)}}{p^0_{(A_{(s+1)}=1|A_{(s+2)}\neq 2)}}, \alpha_{s,4}^{f^1} = \frac{p^0_{(N_{(s+1)}=K_{(s+1)}-1\,|A_{(s+2)}\neq 2)}}{p^{\mathrm{k}-1}_{(A_{(s+1)}=1|\,A_{(s+2)}\neq 2)}}$$

$$\alpha_{s,5}^{e^1} = \frac{p^0_{(N_{(s+1)}=1\,|A_{(s+2)}=2)}}{p^{\mathrm{k}-1}_{(A_{(s+1)}=1|A_{(s+2)}=2)}}, \alpha_{s,5}^{f^1} = \frac{p^0_{(N_2=K_{(s+1)}-1\,|A_{(s+2)}=2)}}{p^0_{(A_{(s+1)}=1|A_{(s+2)}=2)}}$$

$$\alpha_{s,6}^{e^1} = \frac{p^0_{(N_3=1)}}{p^0_{(A_3=1)}}, \alpha_{s,6}^{f^1} = \frac{p^0_{(N_3=K_3-1)}}{p^0_{(A_3=1)}}$$

Else

1) For each sub-network *s,* calculate the total arrival rate to its first queue, and the effective service rate to its third queue from equations (25) and (29)

$$\lambda_s^k = \gamma_s + \frac{\lambda_{s-1}^k(1-p^k_{s-1,A_{,s-1}=2}(0))}{(1-p^k_{s,A_s=2}(0))},$$

$$\widehat{\mu_{s+2}}^k =$$

$$\left(\frac{1}{\mu_{s+2}} + p^k_{s+3,(A_{s+3}=2)}(0)\ \frac{\mu_{s+2}}{\mu_{s+2}+\mu_{s+3}}\ \frac{\lambda^k_{s+1}\left(1-p^k_{s+3,A_{s+3}=2}(0)\right)}{\lambda^k_s\left(1-p^k_{s+2,A_{s+2}=2}(0)\right)}\ \frac{1}{\hat{\mu}_{s+3}}\right)^{-1},$$

where $p^{k-1}_{s,A_s=2}(0)$, $p^{k-1}_{s-1,A_{,s-1}=2}(0)$ are sub-networks s, $s-1$ marginal distributions of queues s, $s-1$ respectively.

2) For each sub-network *s*, calculate the aggregate joint queue-length distribution for time step k:

$$p^k_{s,A=(i,j,l)}(t) = p^{k-1}_{s,A=(i,j,l)}(\delta)e^{t\,Q^k_{AJ_s}}, \quad \forall t \in [0,\delta].$$

where $Q^k_{AJ_s} = f(\gamma,\mu,\alpha^k,B)$ is a 27x27 sparse matrix with nonzero elements described in appendix A. The parameters for the matrix are: $\gamma = [\lambda^k_s, \gamma_{(s+1)}, \gamma_{(s+2)}]$, $\mu = [\mu_s, \mu_{(s+1)}, \widehat{\mu_{(s+2)}}^k]$, $B = [B_1, B_2, B_3, B_4]$ given in Table 2-1,

$\alpha^k = [\alpha_{s,1}^{e^k} \ , \alpha_{s,1}^{f^k}, \alpha_{s,2}^{e^k} \ , \alpha_{s,2}^{f^k} \ , \alpha_{s,3}^{e^k} \ , \ \alpha_{s,3}^{f^k} \ , \ \alpha_{s,4}^{e^k} \ , \ \alpha_{s,4}^{f^k} \ , \ \alpha_{s,5}^{e^k} \ ,$

$\alpha_{s,5}^{f^k} \ , \ \alpha_{s,6}^{e^k} \ , \ \alpha_{s,6}^{f^k} \ .$

3) For each sub-network *s* except the last, where *s* is the index of the sub-network, solve three nonlinear system of equations for the first three blocking scenarios to obtain $\lambda_{s,j}^k \mu_{s,j}^k$ where $j\in$ {1,2,3} is the blocking scenario index. For the last sub-network *s=M-2*, we solve three additional nonlinear systems for $j\in$ {4,5,6} to obtain $\lambda_{s,j}^k \mu_{s,j}^k$.The reason we do this is because the first queue of each sub-network has the most blocking scenarios than any of the other queues in the sub-network, which means capturing the most information on the dependencies between the queues in the sub-network.

Nonlinear system 1: Solve to obtain $\lambda_{s,1}^k, \mu_{s,1}^k$, where s is the index of the sub-network with queue indices $(s, s+1, s+2)$

$$p_{s,(A_s=0|A_{(s+1)}\neq 2)}^{\mathrm{k}}(\delta) - \left(\sum_{m=0}^{K_s} p_{s,(N_s=m|A_{(s+1)}\neq 2)}^{k-1}(\delta)\ d_{N=0}^{m,k}(\delta) \right) = 0$$

$$p_{s,(A_s=2|A_{(s+1)}\neq 2)}^{k}(\delta) - \left(\sum_{m=0}^{K_s} p_{s,(N_s=m|A_{(s+1)}\neq 2)}^{k-1}(\delta)\ \ d_{N=K_s}^{m,k}(\delta) \right) = 0$$

Nonlinear system 2: Solve to obtain $\lambda_{s,2}^k, \mu_{s,2}^k$

$$p_{s,(A_s=0\ |A_2=2,A_{(s+2)}\neq 2)}^{k}(\delta) -$$

$$\left(\sum_{m=0}^{K_s} p_{s,(N_s=m|A_{(s+1)}=2,A_{(s+2)}\neq 2)}^{k-1}\ (\delta)\ d_{N=0}^{m,k}(\delta)\right) = 0$$

$$p_{s,(A_s=2\ |A_{(s+1)}=2,A_{(s+2)}\neq 2)}^{k}(\delta) -$$

$$\left(\sum_{m=0}^{K_s} p_{s,(N_s=m|A_{(s+1)}=2,A_{(s+2)}\neq 2)}^{k-1}\ (\delta) d_{N=K_1}^{m,k}(\delta)\right) = 0$$

Nonlinear system 3: Solve to obtain λ_3^k, μ_3^k

$$p^{k}_{s,(A_s=0\ |A_{(s+1)}=2,A_{(s+2)}=2)}(\delta) -$$

$$\left(\sum_{m=0}^{K_s} p^{k-1}_{s,(N_s=m|A_{(s+1)}=2,A_{(s+2)}=2)}(\delta)\ \ d^{m,k}_{N=0}(\delta)\right) = 0$$

$$p^{k}_{s,(A_s=2\ |A_{(s+1)}=2,A_{(s+2)}=2)}(\delta) -$$

$$\left(\sum_{m=0}^{K_s} p^{k-1}_{s,(N_s=m|A_{(s+1)}=2,A_{(s+2)}\neq 2)}\ (\delta 0) d^{m,}_{N=K_1}(\delta)\right) = 0$$

For *s=M-2* proceed to solve the following as well:

Nonlinear system 4: Solve to obtain $\lambda^{k}_{s,4}, \mu^{k}_{s,4}$

$$p^{k}_{s,(A_{(s+1)}=0\ |A_{(s+2)}\neq 2)}(\delta) -$$

$$\left(\sum_{m=0}^{K_{(s+1)}} p^{k-1}_{s,(N_{(s+1)}=m\ |A_{(s+2)}\neq 2)}(\delta)\ \ d^{m,k}_{N=0}(\delta)\right) = 0$$

$$p^{k}_{s,(A_{(s+1)}=2\ |A_{(s+2)}\neq 2)}(0) -$$

$$\left(\sum_{m=0}^{K_{(s+1)}} p^{k-1}_{s,(N_{(s+1)}=m\ |A_{(s+2)}\neq 2)}\ (0)\ d^{m,k-1}_{N=K_1}(\delta)\right) = 0$$

Nonlinear system 5: Solve to obtain $\lambda^{k}_{s,5}, \mu^{k}_{s,5}$

$$p^{k}_{s,(A_{(s+1)}=0\ |A_{(s+2)}=2)}(\delta) - \left(\sum_{m=0}^{K_{(s+1)}} p^{k-1}_{s,(N_{(s+1)}=m\ |A_{(s+2)}=2)}(\delta)\ \ d^{m,k}_{N=0}(\delta)\right) = 0$$

$$p^{k}_{s,(A_{(s+1)}=2\ |A_{(s+2)}=2)}(0) - \left(\sum_{m=0}^{K_{(s+1)}} p^{k-1}_{s,(N_{(s+1)}=m\ |A_{(s+2)}=2)}\ (\delta)\ d^{m,k}_{N=K_1}(\delta)\right) = 0$$

Nonlinear system 6: Solve to obtain $\lambda^{k}_{s,6}, \mu^{k}_{s,6}$

$$p^{k}_{s,(A_{(s+2)}=0)}(\delta) - \left(\sum_{m=0}^{K_{(s+2)}} p^{k-1}_{s,(N_{(s+2)}=m)}(\delta)\ \ d^{m,k}_{N=0}(\delta)\right) = 0$$

$$p^{k}_{s,(A_{(s+2)}=2)}(\delta) - \left(\sum_{m=0}^{K_{(s+2)}} p^{k-1}_{s,(N_{(s+2)}=m)}\ (\delta)\ d^{m,k}_{N=K_1}(\delta)\right) = 0$$

4) For each sub-network *s,* plug each pair of $\lambda^{k}_{j}, \mu^{k}_{j}$ and previous disaggregate for the blocking scenario as the initial distribution into

Equation (12) to get the disaggregate probabilities of being in disaggregate states *1, K-1*.

Plug $\lambda^{k}_{s,1}, \mu^{k}_{s,1}$ into equation (12) to get

$$p^{\mathrm{k}}_{s,(N_s=1|A_{(s+1)}\neq 2)}(\delta) = \sum_{m=0}^{K_s} p^{k-1}_{s,(N_s=m|A_{(s+1)}\neq 2)}(\delta) d^{m,k}_{N=1}(\delta),$$

$$p^{\mathrm{k}}_{s,(N_s=K_s-1|A_{(s+1)}\neq 2)}(\delta) = \sum_{m=0}^{K_s} p^{k-1}_{s,(N_s=m|A_{(s+1)}\neq 2)}(0) d^{m,k}_{N=K_s-1}(\delta).$$

Plug $\lambda^{k}_{s,2}, \mu^{k}_{s,2}$ into equation (12) to get

$$p^{k}_{s,(N_s=1\;|A_{(s+1)}=2,A_{(s+2)}\neq 2)}(\delta) =$$
$$\textstyle\sum_{m=0}^{K_s} p^{k-1}_{s,(N_s=m|A_{(s+1)}=2,A_{(s+2)}\neq 2)}(\delta) d^{m,k}_{N=1}(\delta),$$
$$p^{k}_{s,(N_s=K_s-1\;|A_{(s+1)}=2,A_{(s+2)}\neq 2)}(\delta) \; =$$
$$\textstyle\sum_{m=0}^{K_s} p^{k-1}_{s,(N_s=m|A_{(s+1)}=2,A_{(s+2)}\neq 2)}(\delta)\; d^{m,k}_{N=K_s-1}(\delta).$$

Plug $\lambda^{k}_{s,3}, \mu^{k}_{s,3}$ into equation (12) to get

$$p^{k}_{s,(N_s=1\;|A_{(s+1)}=2,A_{(s+2)}=2)}(\delta) =$$
$$\textstyle\sum_{m=0}^{K_s} p^{k-1}_{s,(N_s=m|A_{(s+1)}=2,A_{(s+2)}=2)}(\delta) d^{m,k}_{N=1}(\delta),$$
$$p^{k}_{s,(N_s=K_s-1\;|A_{(s+1)}=2,A_{(s+2)}=2)}(\delta) \; =$$
$$\textstyle\sum_{m=0}^{K_s} p^{k-1}_{s,(N_s=m|A_{(s+1)}=2,A_{(s+2)}=2)}(\delta)\; d^{m,k}_{N=K_s-1}(\delta).$$

For *s=M-1*, calculate the following as well:

Plug $\lambda^{k}_{s,4}, \mu^{k}_{s,4}$ into equation (12) to get

$$p^{\mathrm{k}}_{s,(N_{(s+1)}=1|A_{(s+2)}\neq 2)}(\delta) = \textstyle\sum_{m=0}^{K_{s+1}} p^{k-1}_{s,(N_{(s+1)}=m|A_{(s+2)}\neq 2)}(\delta) d^{m,k}_{N=1}(\delta),$$

$p^{\mathrm{k}}_{s,(N_{(s+1)}=K_{(s+1)}-1|A_{(s+2)}\neq 2)}(\delta) =$

$\sum_{m=0}^{K_{s+1}} p^{k-1}_{s,(N_{(s+1)}=m|A_{(s+2)}\neq 2)}(\delta) d^{m,k}_{N=K_{(s+1)}-1}(\delta).$

Plug $\lambda^{k}_{s,5}, \mu^{k}_{s,5}$ into equation (12) to get

$p^{\mathrm{k}}_{s,(N_{(s+1)}=1|A_{(s+2)}=2)}(\delta) = \sum_{m=0}^{K_{s+1}} p^{k-1}_{s,(N_{(s+1)}=m|A_{(s+2)}=2)}(\delta) d^{m,k}_{N=1}(\delta),$

$p^{\mathrm{k}}_{s,(N_{(s+1)}=K_{(s+1)}-1|A_{(s+2)}=2)}(\delta) =$

$\sum_{m=0}^{K_{s+1}} p^{k-1}_{s,(N_{(s+1)}=m|A_{(s+2)}=2)}(\delta) d^{m,k}_{N=K_{(s+1)}-1}(\delta).$

Plug $\lambda^{k}_{s,6}, \mu^{k}_{s,6}$ in equation (12) to get

$$p^{k}_{s,(N_{(s+2)}=1)}(\delta) = \sum_{m=0}^{K_{s+2}} p^{k-1}_{s,(N_{(s+2)}=m)}(\delta) d^{m,k}_{N=1}(\delta),$$

$$p^{k}_{s,(N_{(s+2)}=K_{(s+2)}-1)}(\delta) = \sum_{m=0}^{K_{s+2}} p^{k-1}_{s,(N_{(s+2)}=m)}(\delta) d^{m,k}_{N=K_{(s+2)}-1}(\delta).$$

5) For each sub-network *s* except the last, calculate ${\alpha^{e}_{s,j}}^{k+1}, {\alpha^{f}_{s,j}}^{k+1}$, for $j \in \{1,2,3\}$ for the next time step. For $s = M - 2$, calculate ${\alpha^{e}_{s,j}}^{k+1}, {\alpha^{f}_{s,j}}^{k+1}$, for $j \in \{1,2,3,4,5,6\}$ for the next time step:

$${\alpha^{e}_{s,1}}^{k+1} = \frac{p^{\mathrm{k}}_{s,(N_s=1\ |A_{(s+1)}\neq 2)}(\delta)}{p^{\mathrm{k}}_{s,(A_s=1|A_{(s+1)}\neq 2)}(\delta)}, {\alpha^{f}_{s,1}}^{k+1} = \frac{p^{\mathrm{k}}_{s,(N_s=K_s-1\ |A_{(s+1)}\neq 2)}(\delta)}{p^{\mathrm{k}}_{s,(A_s=1|A_{(s+1)}\neq 2)}(\delta)}$$

$${\alpha^{e}_{s,2}}^{k+1} = \frac{p^{\mathrm{k}}_{s,(N_s=1\ |A_{(s+1)}=2,A_{(s+2)}\neq 2)}(\delta)}{p^{\mathrm{k}}_{s,(A_s=1|\ A_{(s+1)}=2,A_{(s+2)}\neq 2)}(\delta)}, {\alpha^{f}_{s,2}}^{k+1} = \frac{p^{\mathrm{k}}_{s,(N_s=K_s-1\ |A_{(s+1)}=2,A_{(s+2)}\neq 2)}(\delta)}{p^{\mathrm{k}}_{s,(A_s=1|A_{(s+1)}=2\ ,A_{(s+2)}\neq 2)}(\delta)}$$

$${\alpha^{e}_{s,3}}^{k+1} = \frac{p^{\mathrm{k}}_{s,(N_s=1\ |A_{(s+1)}=2,A_{(s+2)}=2)}(\delta)}{p^{\mathrm{k}}_{s,(A_s=1|A_{(s+1)}=2,A_{(s+2)}=2)}(\delta)}, {\alpha^{f}_{s,3}}^{k+1} = \frac{p^{\mathrm{k}}_{s,(N_s=K_s-1\ |A_{(s+1)}=2,A_{(s+2)}=2)}(\delta)}{p^{\mathrm{k}}_{s,(A_s=1|A_{(s+1)}=2\ ,A_{(s+2)}=2)}(\delta)}$$

If $s = M - 2$, calculate the following:

$$\alpha_{s,4}^{e\ \ k+1} = \frac{p^{\mathrm{k}}_{s,(N_{(s+1)}=1\ |A_{(s+2)}\neq 2)}(\delta)}{p^{\mathrm{k}}_{s,(A_{(s+1)}=1|A_{(s+2)}\neq 2)}(\delta)}\ , \alpha_{s,4}^{f\ \ k+1} = \frac{p^{\mathrm{k}}_{s,(N_{(s+1)}=K_{(s+1)}-1\ |A_{(s+2)}\neq 2)}(\delta)}{p^{\mathrm{k}}_{s,(A_{(s+1)}=1|\ A_{(s+2)}\neq 2)}(\delta)}$$

$$\alpha_{s,5}^{e\ \ k+1} = \frac{p^{\mathrm{k}}_{s,(N_{(s+1)}=1\ |A_{(s+2)}=2)}(\delta)}{p^{\mathrm{k}}_{s,(A_{(s+1)}=1|A_{(s+2)}=2)}(\delta)}\ , \alpha_{s,5}^{f\ \ k+1} = \frac{p^{\mathrm{k}}_{s,(N_{(s+1)}=K_{(s+1)}-1\ |A_{(s+2)}=2)}(\delta)}{p^{\mathrm{k}}_{s,(A_{(s+1)}=1|A_{(s+2)}=2)}(\delta)}$$

$$\alpha_{s,6}^{e\ \ k+1} = \frac{p^{\mathrm{k}}_{s,(N_{(s+2)}=1)}(\delta)}{p^{\mathrm{k}}_{s,(A_{(s+2)}=1)}(\delta)}\ , \alpha_{s,6}^{f\ \ k+1} = \frac{p^{\mathrm{k}}_{s,(N_{(s+2)}=K_{(s+2)}-1)}(\delta)}{p^{\mathrm{k}}_{s,(A_{(s+2)}=1)}(\delta)}$$

6) We then infer the rest of the aggregate transition rate probabilities of the second and third queue blocking scenarios from system of Equations (30). For each sub-network *s* except the last, calculate $\alpha_{s,j}^{e\ \ k+1}, \alpha_{s,j}^{f^{k+1}}$, for j ∈ {4,5,6} :

$$\alpha_{s,4}^{e^{k+1}} = \alpha_{s+1,1}^{e^{k+1}}$$
$$\alpha_{s,4}^{f^{k+1}} = \alpha_{s+1,1}^{f^{k+1}}$$

$$\alpha_{s,5}^{e^{k+1}} = p^{k}_{s+1,(A_{s+3}\neq 2)}(\delta)\ \alpha_{s+1,2}^{e^{k+1}} + p^{k}_{i+1,(A_{s+3}=2)}(\delta)\ \alpha_{s+1,3}^{e^{k+1}}$$
$$\alpha_{s,5}^{f^{k+1}} = p^{k}_{s+1,(A_{s+3}\neq 2)}(\delta)\ \alpha_{s+1,2}^{f^{k+1}} + p^{k}_{s+1,(A_{s+3}=2)}(\delta)\ \alpha_{s+1,3}^{f^{k+1}}$$

If $s = M - 3$

$$\alpha_{s,6}^{e^{k+1}} = p^{k}_{s+1,(A_{i+3}\neq 2)}(\delta)\ \alpha_{i+1,4}^{e^{k+1}} + p^{k}_{s+1,(A_{s+3}=2)}(\delta)\ \alpha_{s+1,5}^{e^{k+1}}$$
$$\alpha_{s,6}^{f^{k+1}} = p^{k}_{s+1,(A_{i+3}\neq 2)}(\delta)\ \alpha_{i+1,4}^{f^{k+1}} + p^{k}_{s+1,(A_{s+3}=2)}(\delta)\ \alpha_{s+1,5}^{f^{k+1}}$$

else

$$\begin{aligned}\alpha_{s,6}^{e^{k+1}} &= p^{k}_{s+1,(A_{s+3}\neq 2)}(\delta)\ \alpha_{s+2,1}^{e^{k+1}} \\ &\quad + p^{k}_{s+1,(A_{s+3}=2)}(\delta)\ (p^{k}_{s+2,(A_{s+4}\neq 2)}(\delta)\ \alpha_{s+2,2}^{e^{k+1}} \\ &\quad + p^{k}_{s+2,(A_{s+4}=2)}(\delta)\ \alpha_{s+2,3}^{e^{k+1}})\end{aligned}$$

```
            $\alpha_{s,6}^{f^{k+1}}$
                        $= p_{s+1,(A_{s+3}\neq 2)}^{k}(\delta)\ \alpha_{s+2,1}^{f^{k+1}}$
                        $+ p_{s+1,(A_{s+3}=2)}^{k}(\delta)\ (p_{s+2,(A_{s+4}\neq 2)}^{k}(\delta)\ \alpha_{s+2,2}^{f^{k+1}}$
                        $+ p_{s+2,(A_{s+4}=2)}^{k}(\delta)\ \alpha_{s+2,3}^{f^{k+1}})$
        End
    End
End
```

Chapter 3. Validation

Just as we developed the transient model in three different levels of network sizes, we will validate it for different network sizes. We will first look at a single queue and compare the results we get for the transient queue-length distribution from our model with that we get from the exact model developed by Morse (1958) in equation (12). We will then look at a network of three queues in tandem and compare the transient joint queue-length distribution results we yield from our model with results we get from a discrete event simulator model. Lastly, we will look at tandem networks with different sizes and make comparisons between the transient joint queue-length distributions obtained from our model and those obtained from the discrete event simulator model. Given that our model approximates the transient, we'll present these comparisons at different times including the time at $t = 1{,}10{,}50$. We start with empty queues for all tests in this chapter (i.e., the initial marginal probability of being in aggregate state 0 is set to 1) . The time step used for all the experiments is set to $\delta = 0.1$.

For the discrete event simulator model, we ran 10,000 simulation replications. The distribution results that we got from the event simulator are given for disaggregate states. we derive the aggregate states' distribution from them so we can compare them with results from our model.

We calculate a 95% confidence interval, based on the simulation outputs. We do so by assuming that the sampled probabilities follow a Bernoulli distribution with true value p and sampled value of $\hat{p}$. A %95 confidence interval for p is given from Osorio and Wang (2012) by:

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-p)}{10{,}000 - 1}}\,,\hat{p} + 1.96\sqrt{\frac{\hat{p}(1-p)}{10{,}000 - 1}}\right).$$

The confidence interval is displayed as error bars in figures 3-11 through 3-17.

3.1 Single queue

We consider 10 experiments for testing the transient queue-length distribution, displayed in Table 3-1. The experiments showcase a wide range of values for λ , μ and traffic intensities ρ. The queue capacity is, however, constant with K=10 for all single queue experiments.

The first three plots of figures 3-1 to 3-10 show the results of our model in comparison to results from the exact model given in equation (12). For each experiment, we compare the results at $t = 1,10,50$. By $t = 50$ some of the results from our model reach stationarity. We assume that stationarity is reached when the L2-norm of the change of the distributions for two consecutive iterations is less than 10^{-7}. In the figures below, the blue circles represent the queue-length distributions for each aggregate state from our model, and the red cross represents the queue-length distribution obtained from the exact model. The x-axis in the figures represents the aggregate states (in our case, the aggregate states are: 0,1,2), and the y-axis represents the state probabilities at the time specified in the figures. The last plot of each figure represents the error over time until $t = 50$. The error calculated in these experiments is defined as the difference between the solution from our approximation model, denoted $p_A^{our\ model}$, and the solution from the exact model, denoted p_A^{exact}, for each aggregate state. The errors for aggregate state 0, 1, 2 are defined as $p_{A=0}^{our\ model} - p_{A=0}^{exact}$, $p_{A=1}^{our\ model} - p_{A=1}^{exact}$, $p_{A=2}^{our\ model} - p_{A=2}^{exact}$ respectively.

Experiment	λ	μ	ρ	K
1	0.1	1	0.1	10
2	1	10	0.1	10
3	0.3	1	0.3	10
4	3	10	0.3	10
5	0.7	1	0.7	10
6	7	10	0.7	10
7	0.99	1	0.99	10
8	9.9	10	0.99	10
9	1.2	1	1.2	10
10	12	10	1.2	10

Table 3-1: Experiments to test a single queue

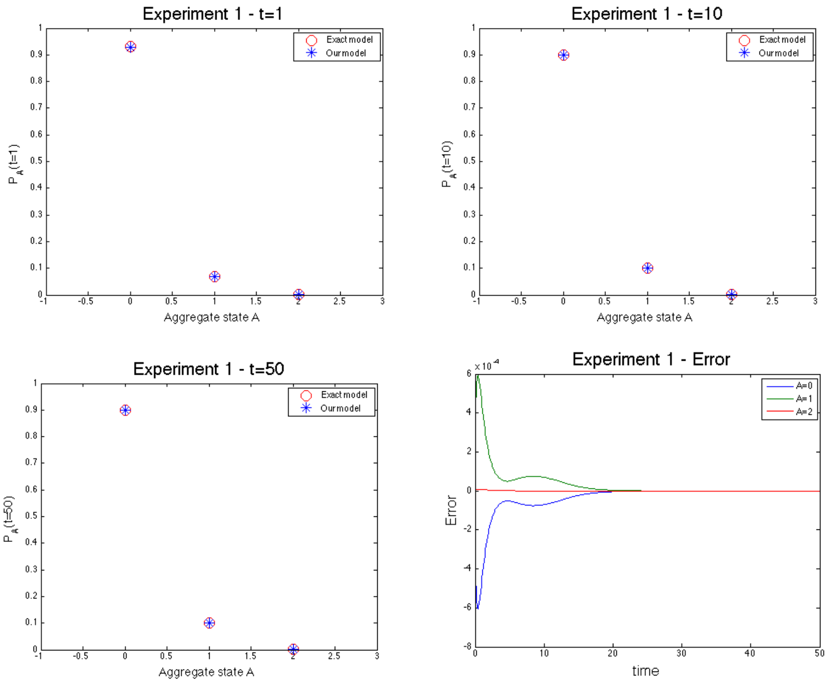


Figure 3-1: Results of experiment 1 at *t* = 1, 10, 50 and the errors for each aggregate state as a function of time

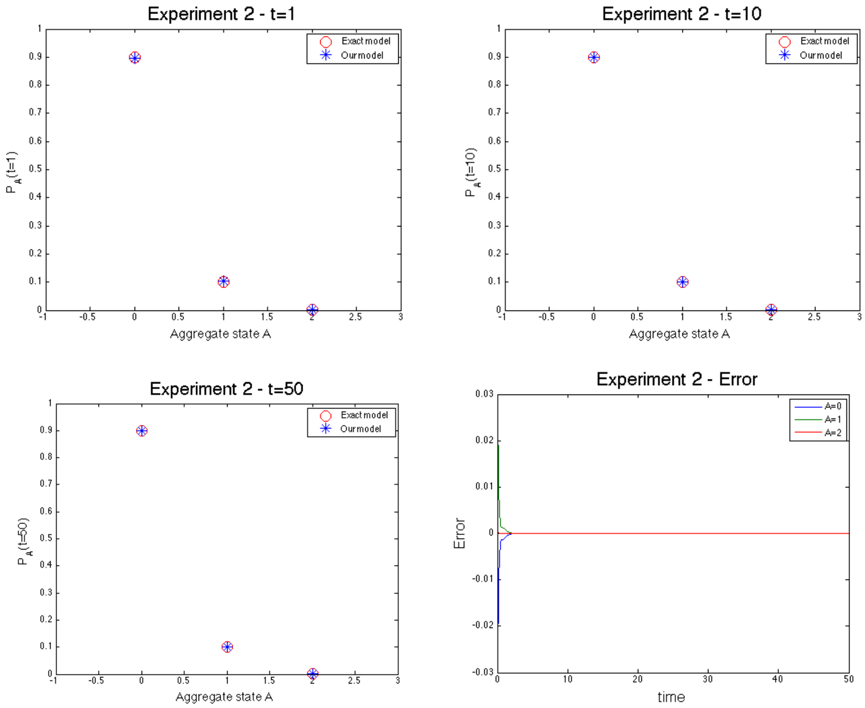


Figure 3-2: Results of experiment 2 at *t* =1, and the errors for each aggregate state as a function of time

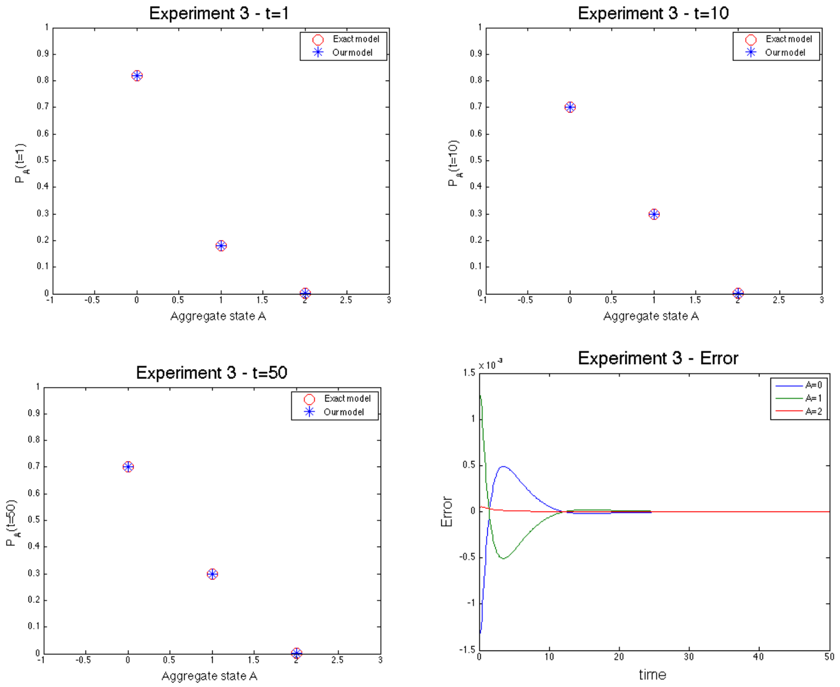


Figure 3-3: Results of experiment 3 at *t* =1,10,50 and the errors for each aggregate state as a function of time

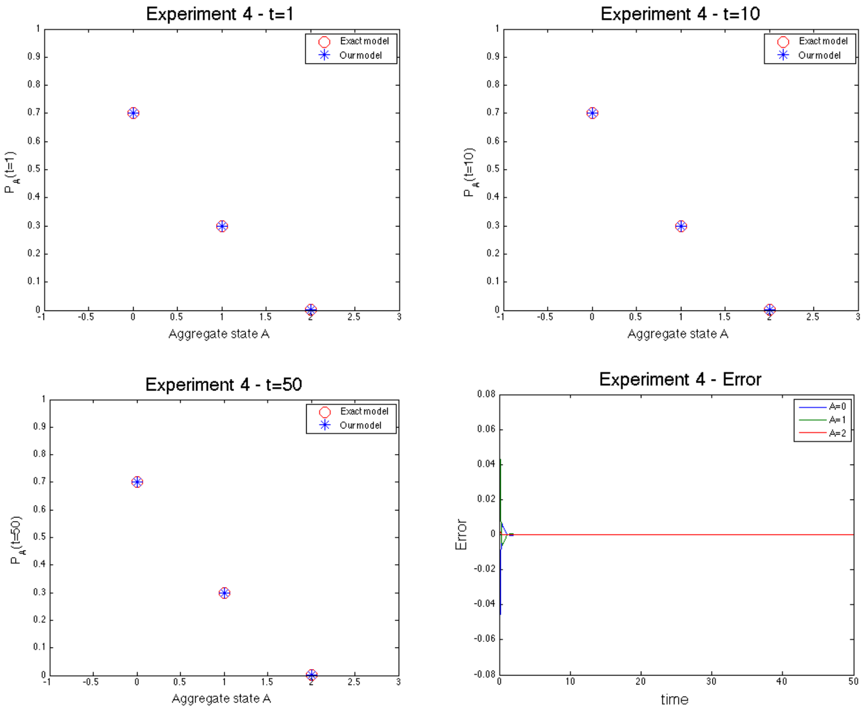


Figure 3-4: Result of experiment 4 at *t*=1,10, 50 and the errors for each aggregate state as a function of time

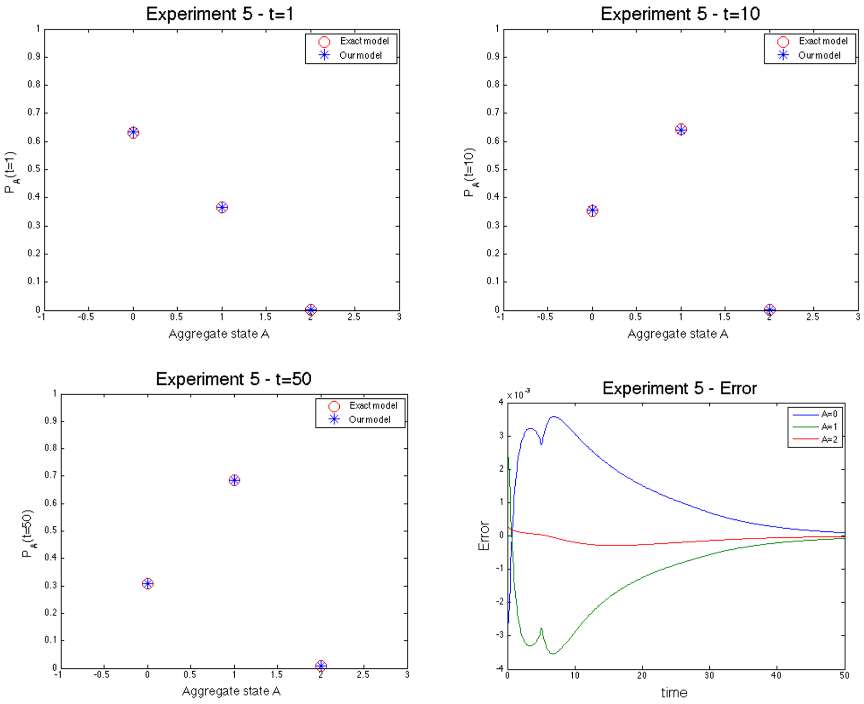


Figure 3-5: Results of experiment 5 at *t* = 1,10, 50 and the errors for each aggregate state as a function of time

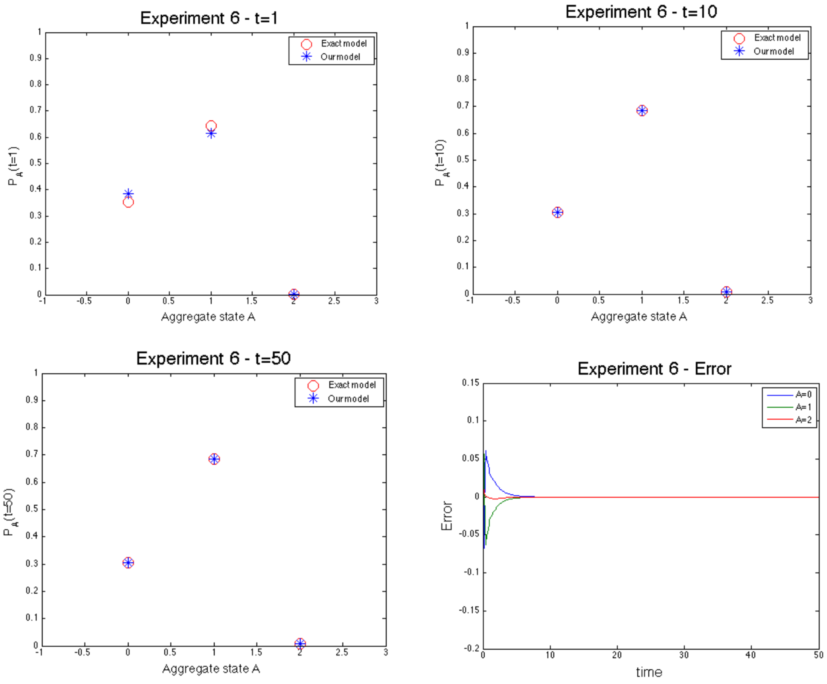


Figure 3-6: Results of experiment 6 at *t*=1,10, 50 and the errors for each aggregate state as a function of time

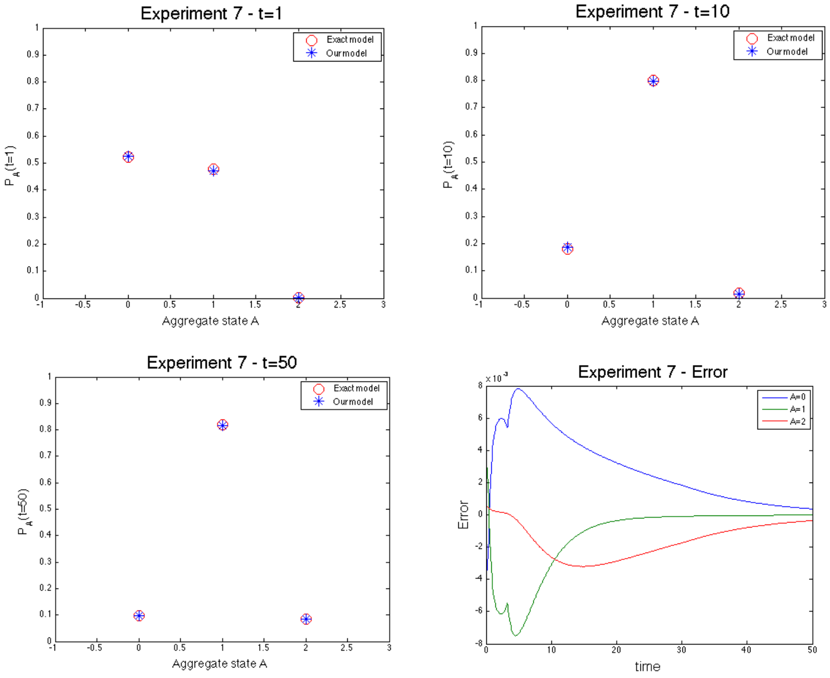


Figure 3-7: Results of experiment 7 at *t* =1,10, 50 and the errors for each aggregate state as a function of time

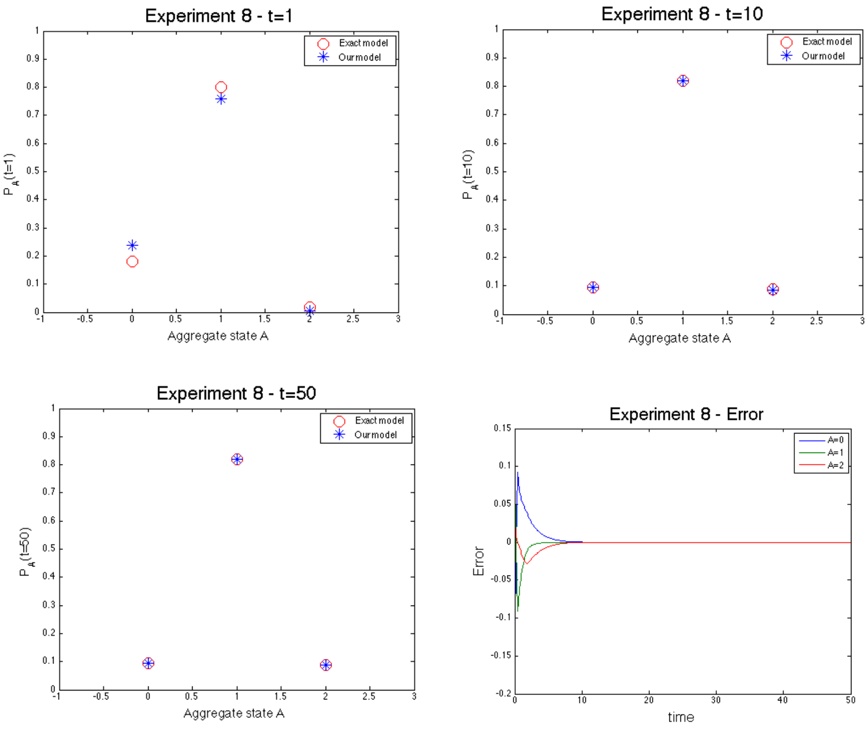


Figure 3-8: Results of experiment 8 at *t*=1,10 and the errors for each aggregate state as a function of time

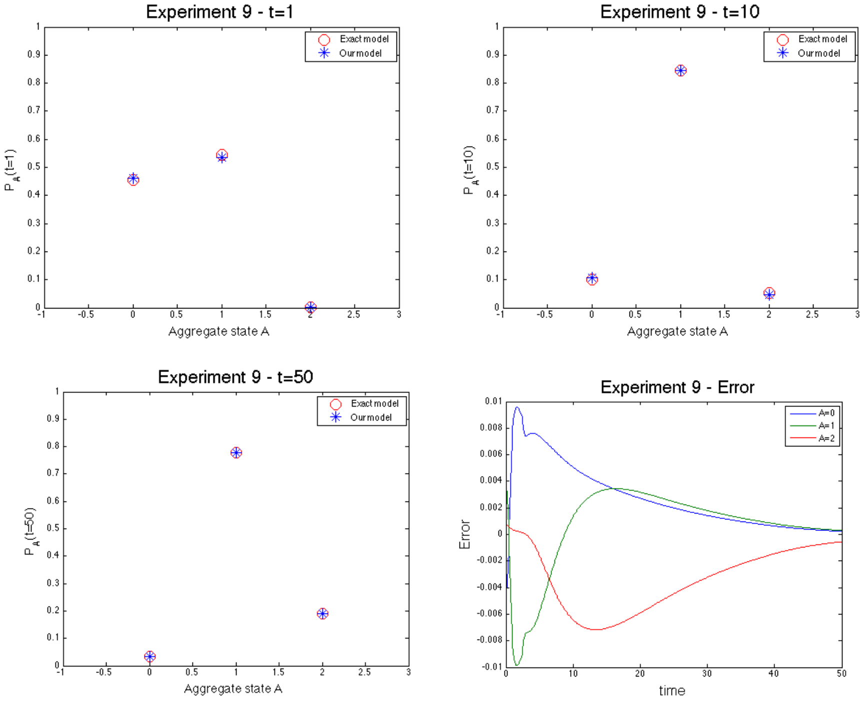


Figure 3-9: Results of experiment 9 at *t*=1,10,50 and the errors for each aggregate state as a function of time

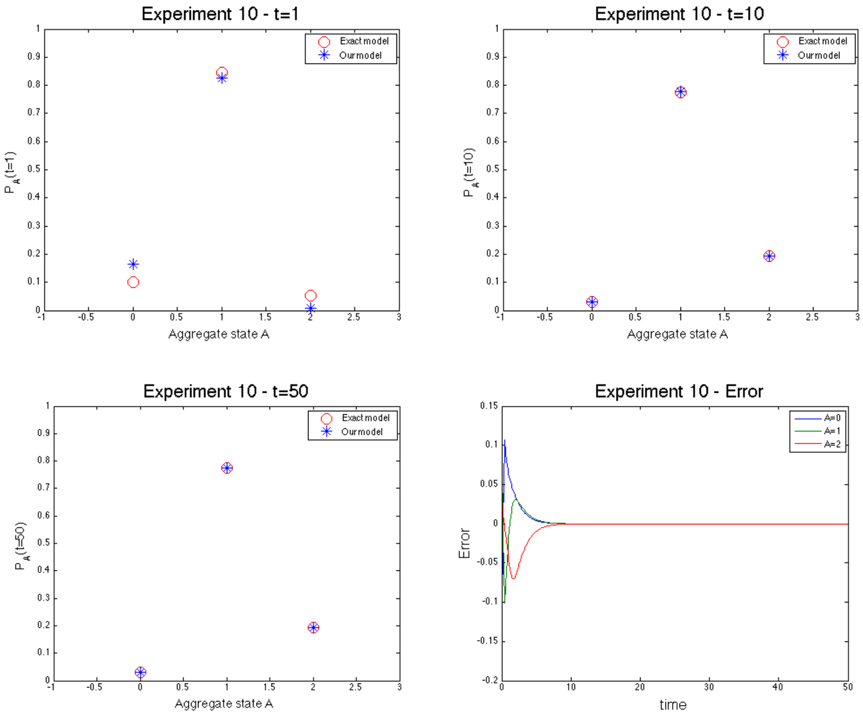


Figure 3-10: Results of experiment 10 at *t*=1,10 and the errors for each aggregate state as a function of time

The graphs show consistent and accurate approximations of the transient queue-length distribution from our model for all experiments. The maximum error that we see is mostly early in the time iteration at $t = 1$. The error decreases exponentially for all experiments until it reaches 10^{-15} for cases when the stationary solution has been reached.

3.2 Three-queue tandem network

Here, we compare results from our three-queue transient joint approximation model with results given by a discrete event simulator. The experiments, displayed in Table 3-2, test a wide range of traffic intensities, some with low traffic intensities ($\rho = 0.3$) and some with high traffic intensities ($\rho = 0.9$). With an empty initial state of the network, we assume an external arrival only to the first queue with the rate $\gamma = 1.8$. For each experiment, we plot the results at time t =1,10,50.

Experiment	γ	μ	ρ	K
1	1.8	[2,2,2]	[0.9, 0.9, 0.9]	[2,2,2]
2	1.8	[2,2,2]	[0.9, 0.9, 0.9]	[5,5,5]
3	1.8	[2,2,2]	[0.9, 0.9, 0.9]	[10,10,10]
4	1.8	[2,4,6]	[0.9, 0.45, 0.3]	[2,2,2]
5	1.8	[2,4,6]	[0.9, 0.45, 0.3]	[5,5,5]
6	1.8	[2,4,6]	[0.9, 0.45, 0.3]	[10,10,10]
7	1.8	[6,4,2]	[0.3, 0.45, 0.9]	[2,2,2]
8	1.8	[6,4,2]	[0.3, 0.45, 0.9]	[5,5,5]
9	1.8	[6,4,2]	[0.3, 0.45, 0.9]	[10,10,10]

Table 3-2: Experiments to test a three-queue network

Each of the figures below displays the aggregate joint queue-length distribution obtained from our model in comparison to the aggregate joint queue-length distribution obtained from the discrete event simulator for each of the 27 aggregate joint states. The blue stars represent the solution from our model and the red circles with error bars represent the solution given from the discrete event simulator. In most of of the experiments in this section, the stationarity is reached by t =50. We define a distribution reaching stationarity when the norm of the difference of the distributions between two consecutive time iterations is less than 10^{-7}.

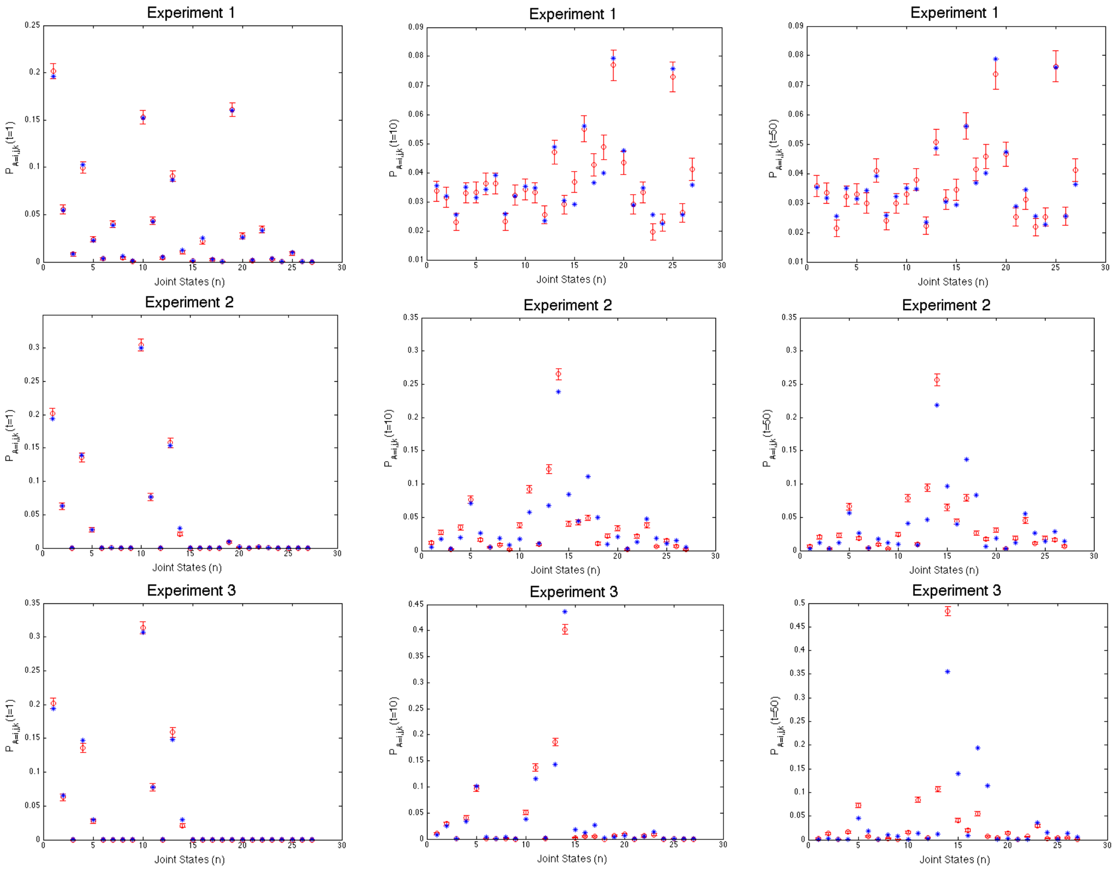


Figure 3-11: Results of experiments 1,2,3 for the 3 joint queue-length distributions with service rate $\mu = [2, 2, 2]$ at t=1,10,50

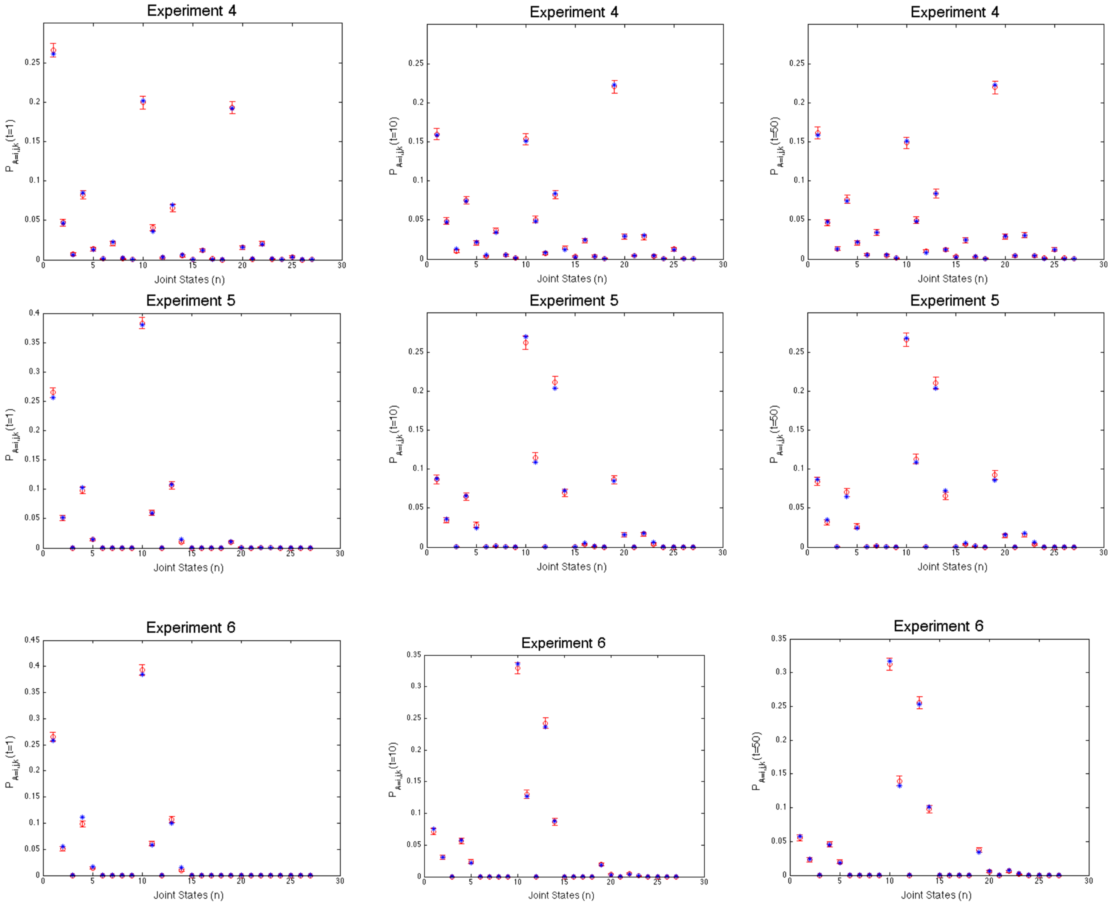


Figure 3-12: Results of experiments 4,5,6 for the 3 joint queue-length distributions with service rate $\mu = [2, 4, 6]$ and t=1,10,50

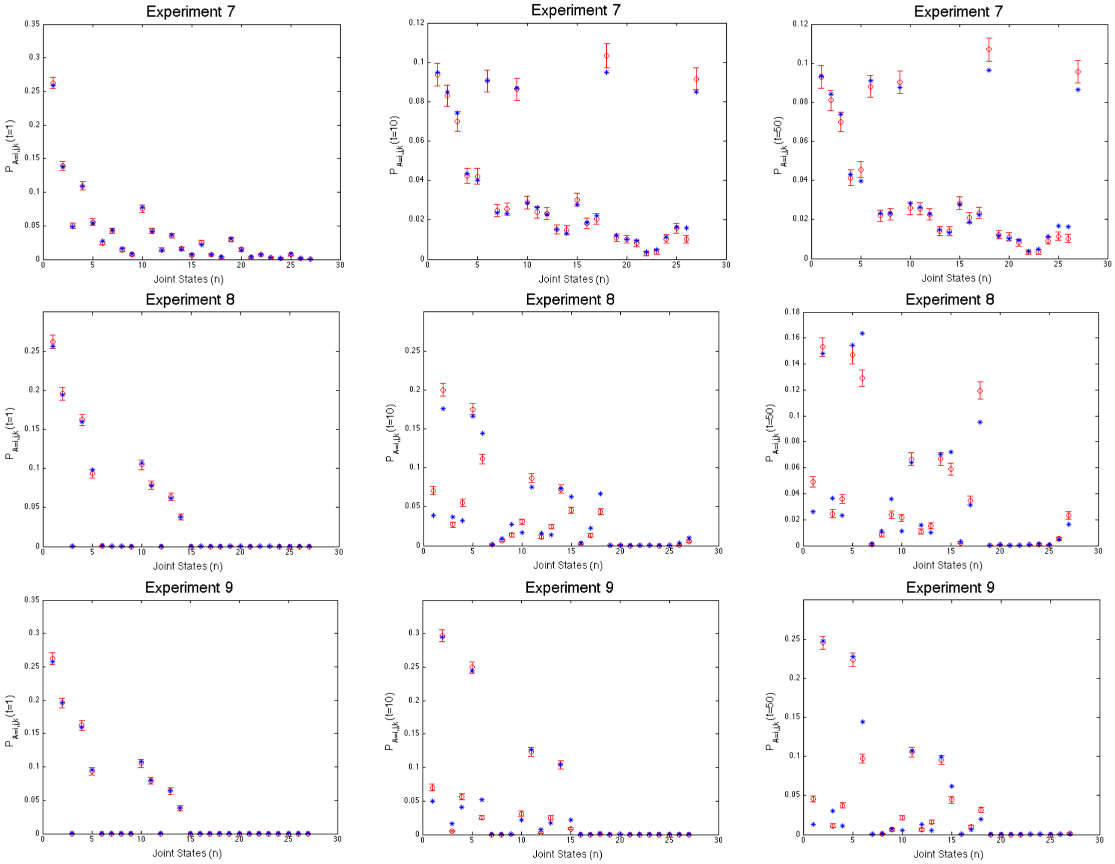


Figure 3-13: Results of experiments 7,8,9 for the 3 joint queue-length distributions with service rate $\mu = [6,4,2]$ and *t*=1,10,50

Experiments 1,4, and 7 show the joint queue-length distributions of queues with capacities 2. If we look at the aggregated queue-length distributions of these experiments, they are equivalent to the disaggregated queue-length distributions because the state space is the same for both the aggregates and disaggregates. The results are, therefore, very accurate at all times.

Experiments 5, and 6 also give precise approximations that are very similar to the simulator results.

Experiments 2,3,8, and 9 start with very accurate results of the joint queue-length distribution similar to those we observe from the discrete even simulator, but because blocking occurs, the joint queue-length distributions become less accurate, but the trends of the distribution from our model and the simulator seems to be similar.

3.3 M-queue tandem network

In this section, we will run experiments for three different network sizes. We start with a fairly small network of five queues in tandem, then a network of eight queues in tandem and conclude with a network of twenty-five queues in tandem. The assumptions mentioned in the beginning of the chapter still applies here where the time step is still $\delta = 0.1$, and all initial marginal distributions for queues are set to 1 for the empty state.

3.3.1 Five queue network

We look at a network of five queues in tandem with parameters displayed in Table 4. The traffic intensity differs for each of the queues. The highest traffic intensity is at the fourth and fifth queues with value $\rho = 0.9$, with blocking most likely to occur. Based on the model we developed in 2.2.3, we get three different overlapping three-queue sub-networks. The corresponding joint queue-length distributions for the three sub-networks are shown in figure 3-14 below. Each row in the figure plots the sub-network solution at three points in time, *t*=1,10,50.

Queue i	γ_i	μ_i	ρ_i	K_i
1	3	10	0.3	25
2	0	10	0.3	10
3	3	10	0.6	25
4	3	10	0.9	10
5	0	10	0.9	25

Table 3-3: Experiment to test a 5-queue tandem network

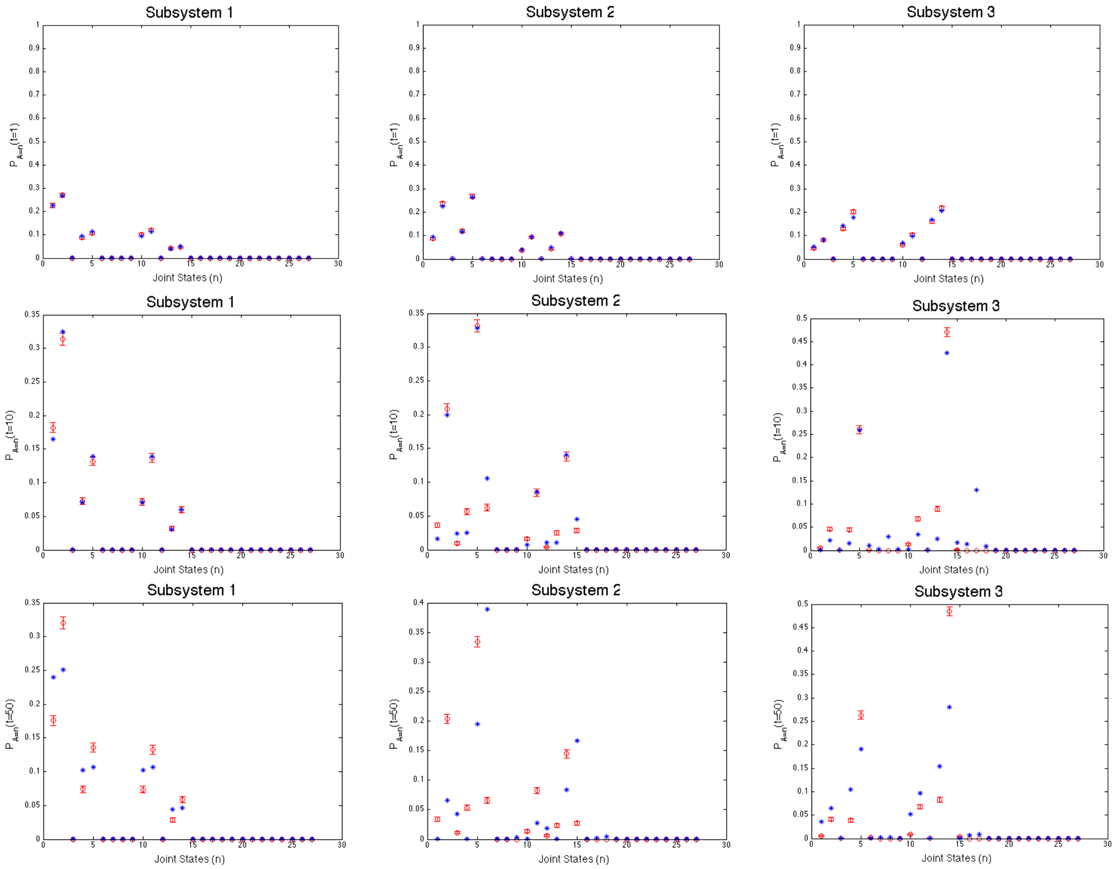


Figure 3-14: Experiment results for the 5 queue joint distribution

The results for the joint queue-length distributions are the most accurate very early on in the time iteration. As time increases, the distribution reached shows similar distribution trends, but not an exact match. At *t*=50, the stationary solution has been reached since the norm of the difference of all the subsystem distributions between two consecutive time iterations is less than 10^{-7}.

3.3.2 Eight queue network

We now look at a network of 8 queues in tandem with 6 overlapping three-queue sub-network. The traffic intensity for this network is similar to the five queue network with blocking most likely to occur at queues 6,7, and 8. The parameters for each queue are displayed in Table 5. Figures 3-15 to 3-17 display the aggregate joint queue-length distribution for each of the 6 subsystems at three points in time *t*=1,10,50.

Queue i	γ_i	μ_i	ρ_i	K_i
1	4	10	0.4	25
2	0	10	0.4	10
3	1	10	0.5	25
4	1	10	0.6	10
5	0	10	0.6	25
6	2	10	0.8	10
7	0	10	0.8	25
8	1	10	0.9	10

Table 3-4: Experiment to test an 8-queue tandem network

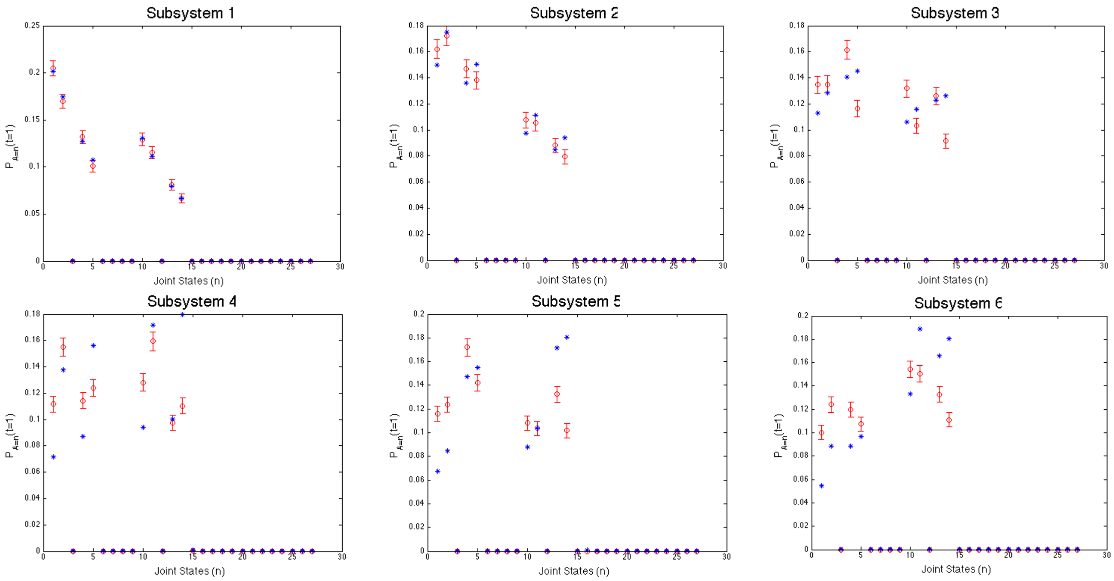


Figure 3-15: Experiment results for the 8-queue joint distribution at t=1

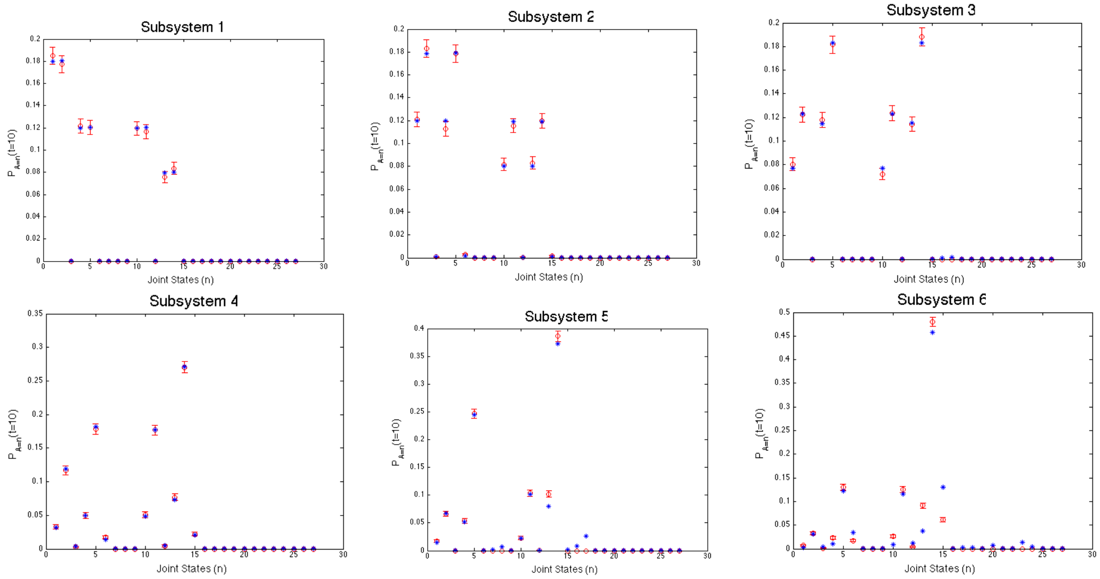


Figure 3-16: Experiment results for the 8-queue joint distribution at t=10

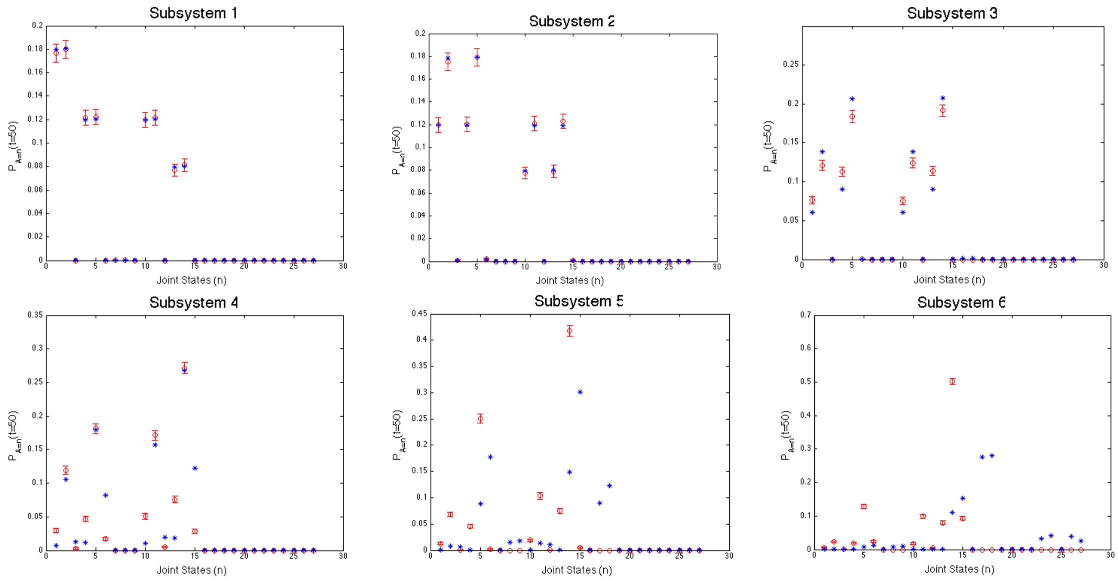


Figure 3-17: Experiment results for the 8-queue joint distribution at t=50

The approximations that we see at *t*=1 are accurate for the first two sub-networks; however, they are not as accurate for the rest of the sub-networks. As time increases, we see that the solutions for the sub-networks that were not very accurate in the beginning become more accurate then it settles on distributions at t=50 with the same trends as the simulator. The stationary distribution is reached at t=50 since the norm of the difference of the distributions for all subsystems is less than 10^{-7}.

3.3.3 Twenty-five queue network

The twenty-five queue network has alternating capacity where odd queues have a capacity of 25 and even queues have a capacity of 10. The service rate for each queue is 10 and external arrival occurs at queues 1,11, and 21 with a rate of 2 with traffic intensity increasing from 0.2 in the first queue to 0.6 on queues 21 onwards. Three histograms of the errors between the results from our model and results from the simulator for all states of the sub-networks at three different times t=1,10,50 are displayed in figures 3-18 to 3-20. For any aggregate joint state probability $A = (i, j, l)$, the error is defined as the difference between the joint distribution obtained from our model and the joint distribution obtained from the simulator, Error= $p_{A=(i,j,l)}^{our\ model} - p_{A=(i,j,l)}^{simulator}$. Our model approximation of the distribution at t=1 is accurate for most states. However, it is more accurate at t=10, 50. In this experiment, by t=50 stationarity was not reached because the norm of the difference of the distribution between two consecutive is not less than 10^{-7}.

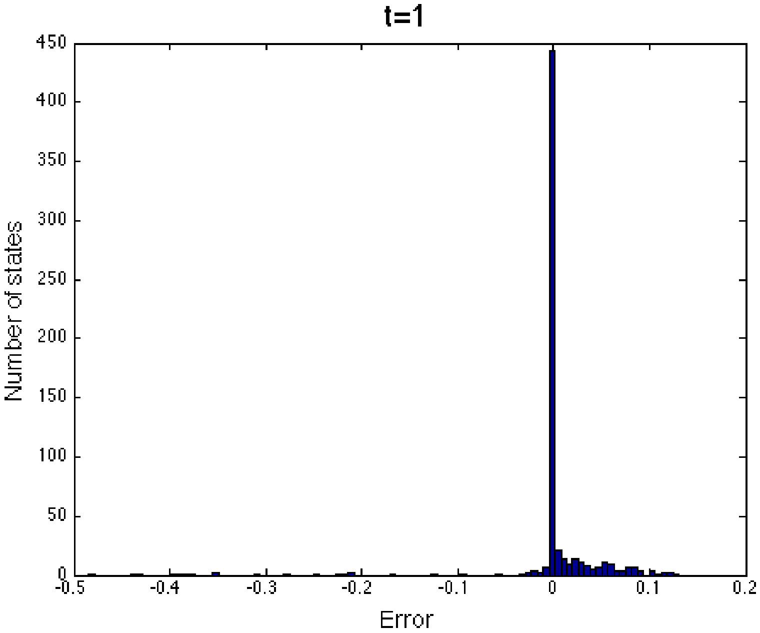


Figure 3-18: Histogram of the errors between the simulated results and the analytical results for each of the 23*26= 621 states of the 25-queue network

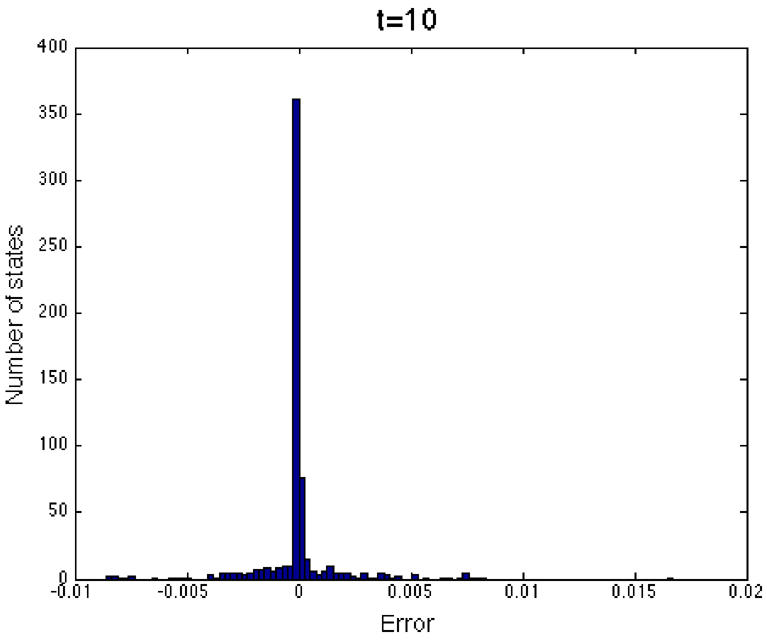


Figure 3-19: Histogram of the errors between the simulated results and the analytical results for each of the 23*26= 621 states of the 25-queue network

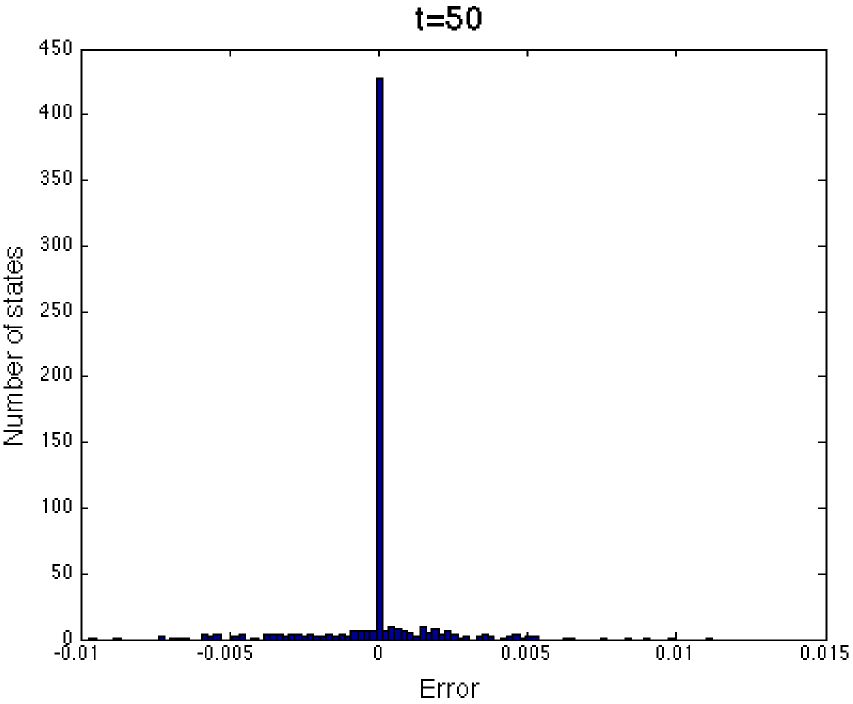


Figure 3-20: Histogram of the errors between the simulated results and the analytical results for each of the 23*26= 621 states of the 25-queue network

Chapter 4. Case Study

Now that we validated our model against an exact model and a discrete event simulator, we are interested in using it to address a traditional urban traffic signal control problem. Our goal for this chapter is to investigate the added value of accounting for the transient joint queue-length distribution for different network demand scenarios. We evaluate our proposed analytically approximated transient joint model by comparing it to an analytically approximated stationary joint model derived by Osorio and Wang (2012).

We model the urban road network as a finite-capacity queueing network by following the approach presented in Osorio and Bierlaire (2009b). Each lane in the model is presented as a queue, and the flow capacities of the lanes correspond to the service rates of the queues.

A microscopic traffic simulation model implemented in AIMSUN, version 6.1, evaluates the performance of the signal plans that are measured by both our model and the stationary joint model. We ran 50 simulation replications, each for an hour with a warm-up period of fifteen minutes. For each replication, we obtained an average trip travel time. The cumulative distribution functions obtained from the 50 replications for both our model and the stationary joint model are then compared.

4.1 Network

We consider the same urban road network studied in Osorio and Wang (2012). The network, displayed in Figure 4-1, consists of 20 single lanes, with 4 intersections, each with 2 endogenous phases. All west-bound links of the main artery are modeled jointly, as are all east-bound links. All cross streets (north-bound and south-bound) are modeled independently. External arrivals and external departures only occur at the boundaries of the network where the blue circles are. Customers can travel along a single direction without making any turns in the network.

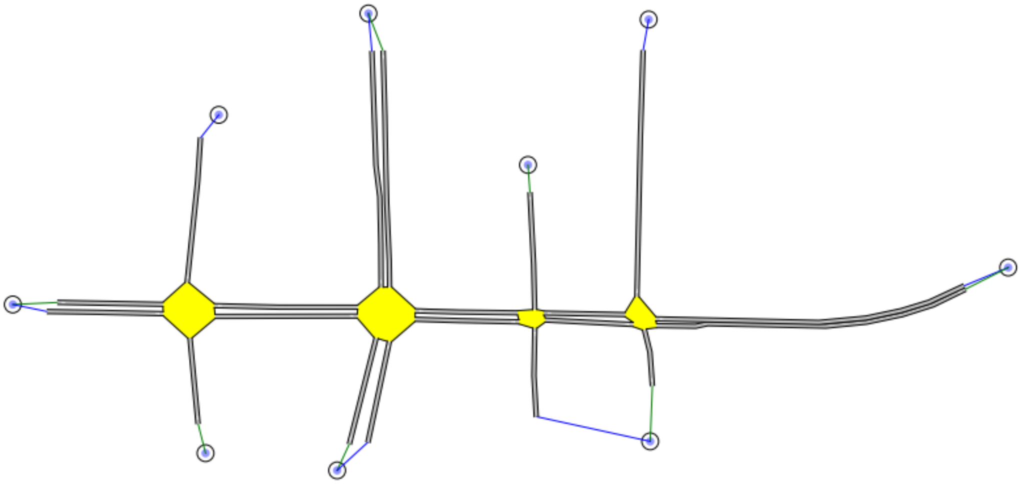


Figure 4-1: Urban road network of single roads for the case study (Osorio and Wang, 2012)

We consider two demand scenarios in the network: medium demand and high demand scenarios. For the medium demand scenario, the east-bound and west-bound demands are 700 vehicles per hour, and the demands for cross streets differ. For the high demand scenario, the east-bound and west-bound demands increase to 900 vehicles per hour. Details on the exact demand for each scenario are displayed in Table 4-1.

Demand scenario	West-bound	East-bound	South-bound-1	South-bound-2	North-bound-2	North-bound-3	South-bound-4
Medium	700	700	100	600	600	100	100
high	900	900	100	600	600	200	200

Table 4-1: Demand in vehicles per hour for the medium and high demand scenarios

4.2 Problem formulation

We follow the signal control problem formulated in details in Osorio and Bierlaire (2009b). We briefly describe it here. We consider a fixed-time control strategy for a specific time duration T with time interval $[(t_0, t_0 + T]$ and initial queue-length distribution of the each lane p_l^0, which are exogenous problem parameters. The strategy uses information on the transient queue-length distribution throughout the entire time interval of interest, $[(t_0, t_0 + T])$, to derive a fixed signal plan. The signal plans are calculated offline and signal plans for multiple intersections are determined jointly. The normalized green times of phases of the different intersections are the decision variables. All other traditional control variables like the cycle time, offsets and stage structure are assumed fixed.

We introduce the following notation for the problem:

b_i available cycle ratio of intersection *i;*

s saturation flow rate [veh/h];

$x(j)$ green split of phase *j*;

x_L vector of minimal green splits;

$\mathcal{I}$ set of intersection indices;

$\mathcal{L}$ set of indices of the signalized lanes;

$\mathcal{P}_I(i)$ set of phase indices of intersection *i*;

$\mathcal{P}_L(l)$ set of phase indices of lane *l*;

t_0 starting time of the interval of interest;

T total duration of time interval of interest;

p_l^0 initial marginal queue-length distributions of lane *l*;

$y(T)$ vector of time-dependent endogenous queueing variables (e.g., disaggregation probabilities);

u vector of exogenous queueing parameters (e.g., external arrival rate, space capacities).

δ time step length

The problem is formulated as follows:

$$\min_{\mathrm{x}} A(T, x, y(T), u) \tag{31}$$

subject to

$$\sum_{j \in \mathcal{P}_I(i)} x_j = b_i, \forall\, i \in \mathcal{I}\,, \tag{32}$$

$$\mu_l - \sum_{j \in \mathcal{P}_L(l)} x_j s = 0\,, \forall\, l \in \mathcal{L}, \tag{33}$$

$$h(T, y(T), u, x) = 0 \tag{34}$$

$$y(T) \geq 0, x \geq x_L. \tag{35}$$

The decision vector x consists of the green times for each phase. Equation (32) in the constraints ensures that the available cycle time of each intersection are distributed among the phases of the intersection. Equation (33) in the constraints relates the service rate of signalized queue μ_l to the saturation flow s (set to 1800 vehicles per hour) and to the green split of its phases, x_j. Equation (34) represents the equations for the time-dependent queueing model that if solved yields the transient queue-length distribution of the network. The queueing model, h, depends on a time-dependent vector of endogenous variables $y(T)$, and a set of exogenous queue parameters u as well as the decision vector x. In Equation (35) of the constraints, the endogenous queue variables are subject to positivity constraints and green splits x have lower bounds which are set to 4 seconds here (following the transportation norms VSS (1992)). The objective function $A(T, x, y(T), u)$ in Equation (31) represents the expected trip travel time during the period $[t_0, t_0 + T]$ which depends on $y(T)$, u and the vector of green splits for each phase x.

If we denote $E\left[N_i^k\right]$ as the expected number of vehicles in queue i at the end of time interval k, and $p_{(N_i<K_i)}^k(\delta)$ as the probability of link *i* being full at the end of time interval k, then the expected travel time during discrete time interval k, denoted $A^k(T,x,y(T),u)$, can be approximated with Little's law (Little, 2011, 1961):

$$A^k(T,x,y(T),u) = \frac{\sum_i E\left[N_i^k\right]}{\sum_i \gamma_i p_{(N_i<K_i)}^k(\delta)},$$

(36)

where the summation of $E\left[N_i^k\right]$ considers all queues in the network (queues are indexed by *i)*. Additionally,$p_{(N_i<K_i)}^k(\delta)$ can be calculated from the marginal aggregate queue-length distribution during the end of time interval k from the following equation: $p_{(N_i<K_i)}^k$ (δ)= $(1 - p_{(N_i=K_i)}^k(\delta))$.

Lastly, the expected travel time during the entire simulation period $[t_0, t_0 + T]$ can be approximated as:

$$A(T,x,y(T),u) = \frac{\sum_{k=\frac{t_0}{\delta}}^{\frac{T+t_0}{\delta}} A^k(T,x,y(T),u)}{I},$$

(37)

where I is the total number of time intervals, and is equal to $\frac{T+t_0}{\delta} - \frac{t_0}{\delta}$, assuming both $\frac{t_0}{\delta}$ and $\frac{T+t_0}{\delta}$ are integers.

The derivation of $E[N_i^k]$, on the other hand, is calculated from the disaggregate queue-length distribution of queue i during end of time interval k. The disaggregate distribution for an individually modeled queue is given from solving the nonlinear system of equations given in Equations (15) and (16) that yield the rates λ, μ, at time interval k and plugging then into equation (12) to get all disaggregate state probabilities of queue *i* during that time interval. Additionally, the marginal disaggregate queue-length distribution at end of time interval *k* for a jointly modeled queue *i* is given from analyzing the joint aggregate distribution of sub-network *i* during the end of time interval *k,* which is done by first calculating the marginal aggregate distribution of queue i, then disaggregating it similarly as the individually modeled queue approach. To calculate the expected number of vehicles in all queues of the network at end of time interval k, we use the following equation:

$$E[N_i^k] = \sum_{n=0}^{K_i} n\, p_{(N_i=n)}^k(\delta),$$

(38)

where $p_{(N_i=n)}^k(\delta)$ is the probability that queue i is in disaggregate state n during end of time interval k.

The difference between our formulation of the signal control problem and the stationary model's formulation of the signal control problem is that the stationary formulation does not depend on time for any of the parameters above and solves for the signal plans based on stationary network information. The stationary formulation also includes a queueing model constraint that is not time-dependent, $h(y; u) = 0$, that depends on the endogenous parameters y as well as the exogenous parameters u. For more details on the stationary model formulation and implementation details of the signal control problem, we refer the reader to chapter 4 of Osorio and Carter (2012).

4.3 Implementation Notes

The case study network (Figure 4-1) is made up of 20 single-lane roads that are modeled as follows: two sets of five-queue networks modeled jointly, and 10 queues, which are modeled individually (not part of a joint network).

We assume empty initial queues for all queues in the network. The time at which we calculate the fixed signal plans is at t=seventy-five minutes (an hour and fifteen minutes which, in the simulation model, is decomposed as a fifteen minute warm-up period and one hour of further simulation). The time step used when calculating the transient queue-length distribution and signal plans is set to $\delta = 0.1$.

The initial signal plans for the 8 phase variables (2 per intersection) used in our model is the optimal plan that we get from a marginal model discussed in Osorio and Wang (2012), the remaining endogenous variables are obtained by calculating the transient queue-length distribution of the jointly modeled queues, as well as the individually modeled queues. The set of variables is used as an initial feasible point for the signal control problem which is then solved using the "active-set" algorithm of the *fmincon* solver of Matlab with constraint and function tolerance of 10^{-6} and 10^{-3}, respectively.

The effective service rates $\hat{\mu}$ are calculated from Equation (26) from the exogenous service rates and the transient joint queue-length distribution at time *t*. The arrival rates λ are calculated from Equation (25) from the external arrival rates and transient joint queue-length distribution at time *t*.

In the signal control problem, we implement the expected number of vehicles, $E\left[N_i^k\right]$, for each queue i in the network during end of time interval k, from the marginal disaggregate probabilities obtained from the transient models.

4.4 Results

4.3.1 Medium demand scenario

We present the cumulative distribution function (cdf) results for the average travel time, displayed in Figure 4-2, for signal plans solved using our model and the stationary joint model. We can see that the signal plan results from our method perform better than the joint stationary model because the cdf from our model is to the left of that from the stationary joint model.

We ran a paired t-test at a 99% confidence level to test the hypothesis that the expected travel time derived from our model is equal to that derived from the stationary joint model for this scenario. The mean of the paired difference, denoted $\bar{X}$, is approximately

0.0768 minutes. The standard deviation, denoted $\hat{s}$, is approximately 0.0243 minutes. For the 50 observations, the paired t-test is given by Hogg and Tanis (2006, p.486):

$$t = \sqrt{50}\frac{\bar{X}}{\hat{s}}$$

Hence, the test statistics of this experiment is 22.315. The null hypothesis is rejected because the critical value, $t_{0.01}(49) = 2.405$ is less than the value of the test statistic for this experiment.

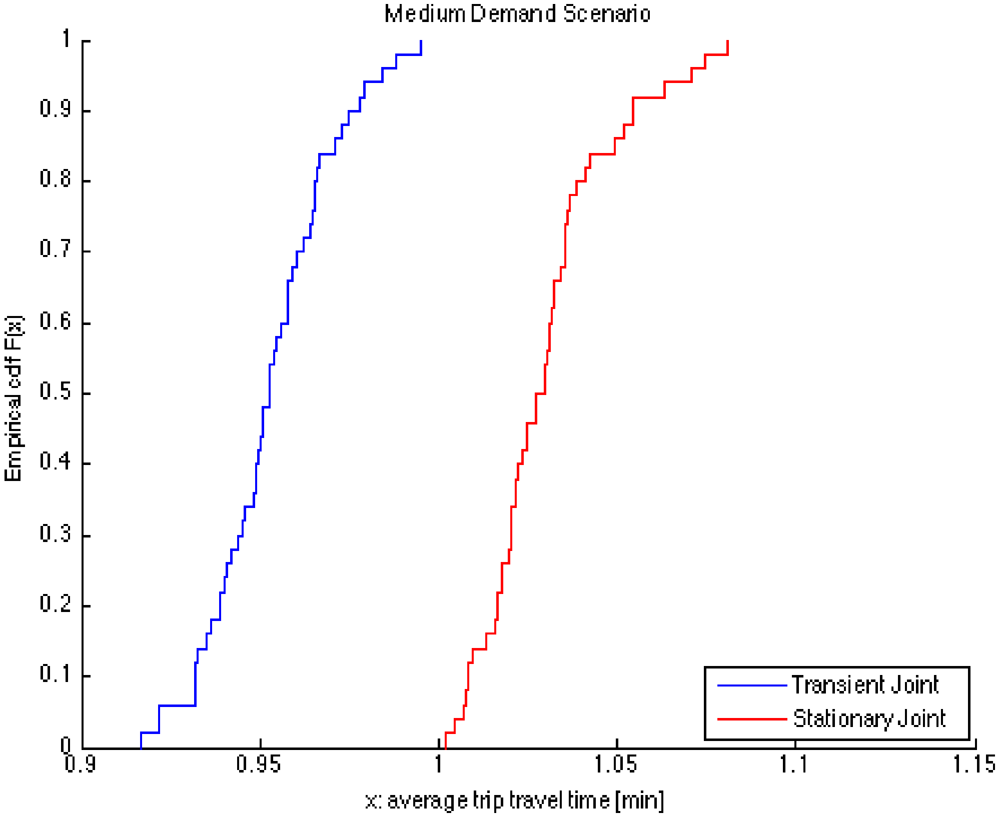


Figure 4-2: CDF's of the average trip travel time for the medium demand test

4.3.2 High demand scenario

As we did with the medium demand scenario, we present the cdf results of the average travel time, displayed in figure 4-2, for signal plans solved using our model and the stationary joint model. Similarly, we can see that the signal plan results from our method perform better than the joint stationary model.

We ran the same paired t-test as the medium scenario, at a 99% confidence level, to test the hypothesis that the expected travel time derived from our model is equal to that derived from the stationary joint model. The mean of the paired difference is approximately 0.0771 minutes. The standard deviation is approximately .0286 minutes. For the 50 observations, the test statistics of this experiment is 19.0564. The null hypothesis is rejected because the critical value, $t_{0.01}(49) = 2.405$ is less than the value of the test statistic for this experiment.

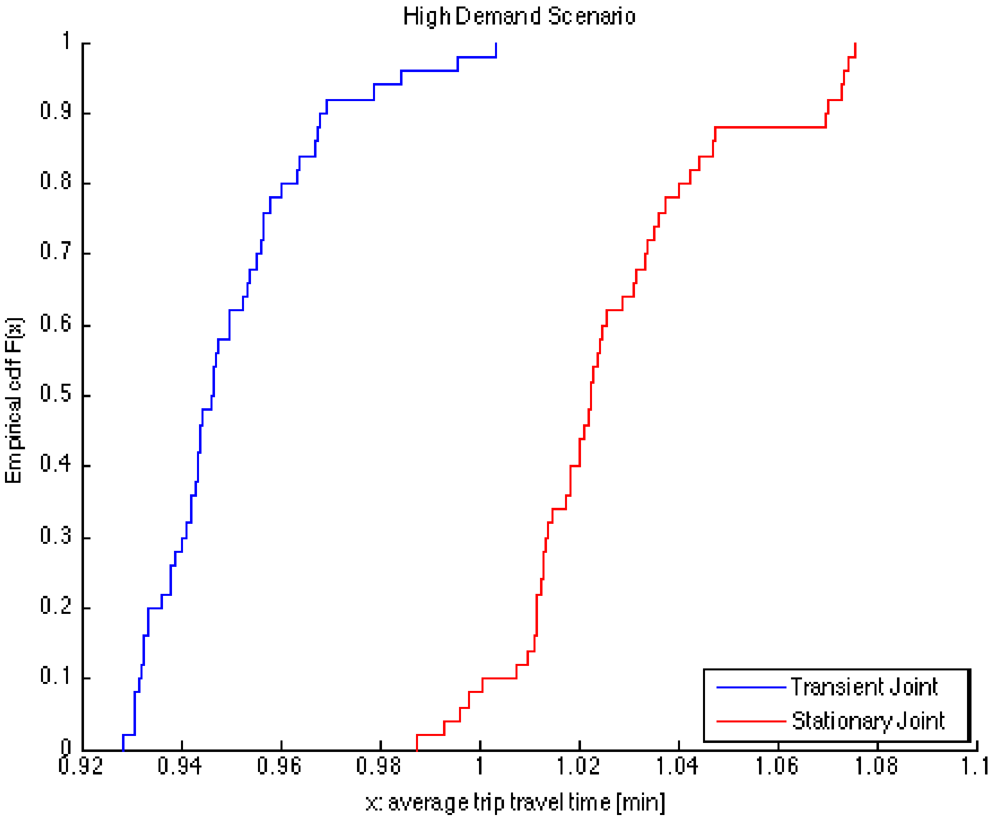


Figure 4-2: CDF's of the average trip travel time for the high demand test

Chapter 5. Conclusions

In this research work, we derived analytical approximation models for the transient queue-length distribution of a single M/M/1/K queue as well as the transient joint queue-length distribution for a tandem network of three M/M/1/K queues in tandem. Both approximations were used to derive the transient joint queue-length distribution of any tandem network size by decomposing the network into overlapping 3-queue sub-networks. The model does not ensure consistency between the marginal queue-length distributions of overlapping queues, however, the model ensures consistency between the aggregate transition rate probabilities of the same queues in different sub-network. The results, in most cases, show accuracy between our model and results obtained from a discrete event simulator. For cases when blocking occurs, we observe similar queue-length distribution trends between results from our model and results from the discrete event simulator. In addition, accounting for the transient queue-length distribution of a network instead of the stationary queue-length distribution of the same network showed better average travel trip times when addressing a signal control problem for both medium and high demand network scenarios.

Appendix A: Transition rate matrix for the three-queue tandem network

Matrix index	From state	To state	Transition rate
(1,2)	{0,0,0}	{0,0,1}	γ_3
(1,4)	{0,0,0}	{0,1,0}	γ_2
(1,10)	{0,0,0}	{1,0,0}	γ_1
(2,1)	{0,0,1}	{0,0,0}	$\mu_3\alpha_6^e$
(2,3)	{0,0,1}	{0,0,2}	$\gamma_3\alpha_6^f$
(2,5)	{0,0,1}	{0,1,1}	γ_2
(2,11)	{0,0,1}	{1,0,1}	γ_1
(3,2)	{0,0,2}	{0,0,1}	μ_3
(3,6)	{0,0,2}	{0,1,2}	γ_2
(3,12)	{0,0,2}	{1,0,2}	γ_1
(4,2)	{0,1,0}	{0,0,1}	$\mu_2\alpha_4^e$
(4,5)	{0,1,0}	{0,1,1}	$\mu_2(1-\alpha_4^e)$
(4,7)	{0,1,0}	{0,2,0}	$\gamma_2\alpha_4^f$
(4,13)	{0,1,0}	{1,1,0}	γ_1
(5,2)	{0,1,1}	{0,0,1}	$\mu_2\alpha_4^e(1-\alpha_6^f)$
(5,3)	{0,1,1}	{0,0,2}	$\mu_2\alpha_4^e\alpha_6^f$
(5,4)	{0,1,1}	{0,1,0}	$\mu_3\alpha_6^e$
(5,6)	{0,1,1}	{0,1,2}	$\mu_2(1-\alpha_4^e)\alpha_6^f$
(5,8)	{0,1,1}	{0,2,1}	$\gamma_2\alpha_4^f$
(5,14)	{0,1,1}	{1,1,1}	γ_1
(6,3)	{0,1,2}	{0,0,2}	$\mu_3(1-\alpha_5^e)B_1$
(6,5)	{0,1,2}	{0,1,1}	$\mu_3(1-B_1)$
(6,9)	{0,1,2}	{0,2,2}	$\gamma_2\alpha_5^f$
(6,15)	{0,1,2}	{1,1,2}	γ_1
(7,5)	{0,2,0}	{0,1,1}	μ_2
(7,8)	{0,2,0}	{0,2,1}	γ_3
(7,16)	{0,2,0}	{1,2,0}	γ_1
(8,5)	{0,2,1}	{0,1,1}	$\mu_2(1-\alpha_6^f)$
(8,6)	{0,2,1}	{0,1,2}	$\mu_2\alpha_6^f$
(8,7)	{0,2,1}	{0,2,0}	$\mu_3\alpha_6^f$
(8,9)	{0,2,1}	{0,2,2}	$\gamma_3\alpha_6^f$
(8,17)	{0,2,1}	{1,2,1}	γ_1
(9,6)	{0,2,2}	{0,1,2}	$\mu_3(1-B_1)$
(9,8)	{0,2,2}	{0,2,1}	μ_3
(9,18)	{0,2,2}	{1,2,2}	γ_1

(10,4)	{1,0,0}	{0,1,0}	$\mu_1\alpha_1^e$
(10,11)	{1,0,0}	{1,0,1}	γ_3
(10,13)	{1,0,0}	{1,1,0}	$\mu_1(1-\alpha_1^e)$
(10,19)	{1,0,0}	{2,0,0}	$\gamma_1\ \alpha_1^f$
(11,5)	{1,0,1}	{0,1,1}	$\mu_1\alpha_1^e$
(11,10)	{1,0,1}	{1,0,0}	$\mu_3\alpha_6^e$
(11,12)	{1,0,1}	{1,0,2}	$\gamma_3\alpha_6^f$
(11,14)	{1,0,1}	{1,1,1}	$\mu_1(1-\alpha_1^e)$
(11,20)	{1,0,1}	{2,0,1}	$\gamma_1\ \alpha_1^f$
(12,6)	{1,0,2}	{0,1,2}	$\mu_1\alpha_1^e$
(12,11)	{1,0,2}	{1,0,1}	μ_3
(12,15)	{1,0,2}	{1,1,2}	$\mu_1(1-\alpha_1^e)$
(12,21)	{1,0,2}	{2,0,2}	$\gamma_1\ \alpha_1^f$
(13,4)	{1,1,0}	{0,1,0}	$\mu_1\alpha_1^e(1-\alpha_4^f)$
(13,7)	{1,1,0}	{0,2,0}	$\mu_1\alpha_1^e\alpha_4^f$
(13,11)	{1,1,0}	{1,0,1}	$\mu_2\alpha_4^e$
(13,14)	{1,1,0}	{1,1,1}	$\mu_2(1-\alpha_4^e)$
(13,16)	{1,1,0}	{1,2,0}	$\mu_1\alpha_4^f(1-\alpha_1^e)$
(13,22)	{1,1,0}	{2,1,0}	$\gamma_1\ \alpha_1^f$
(14,5)	{1,1,1}	{0,1,1}	$\mu_1\alpha_1^e(1-\alpha_4^f)$
(14,8)	{1,1,1}	{0,2,1}	$\mu_1\alpha_1^e\alpha_4^f$
(14,11)	{1,1,1}	{1,0,1}	$\mu_2\alpha_4^e(1-\alpha_6^f)$
(14,12)	{1,1,1}	{1,0,2}	$\mu_2\alpha_4^e\alpha_6^f$
(14,13)	{1,1,1}	{1,1,0}	$\mu_3\alpha_6^e$
(14,15)	{1,1,1}	{1,1,2}	$\mu_2(1-\alpha_4^e)\alpha_6^f$
(14,17)	{1,1,1}	{1,2,1}	$\mu_1\alpha_4^f(1-\alpha_1^e)$
(14,23)	{1,1,1}	{2,1,1}	$\gamma_1\ \alpha_1^f$
(15,6)	{1,1,2}	{0,1,2}	$\mu_1\alpha_1^e(1-\alpha_5^f)$
(15,9)	{1,1,2}	{0,2,2}	$\mu_1\alpha_1^e\alpha_5^f$
(15,12)	{1,1,2}	{1,0,2}	$\mu_3B_2\alpha_5^e$
(15,14)	{1,1,2}	{1,1,1}	$\mu_3(1-B_2)$
(15,18)	{1,1,2}	{1,2,2}	$\mu_1\alpha_5^f(1-\alpha_1^e)$
(15,24)	{1,1,2}	{2,1,2}	$\gamma_1\ \alpha_1^f$
(16,8)	{1,2,0}	{0,2,1}	$\mu_2B_1\alpha_1^f$
(16,14)	{1,2,0}	{1,1,1}	$\mu_2(1-B_1)$
(16,17)	{1,2,0}	{1,2,1}	$\mu_2B_1(1-\alpha_1^f)$
(16,25)	{1,2,0}	{2,2,0}	$\gamma_1\ \alpha_2^f$
(17,8)	{1,2,1}	{0,2,1}	$\mu_2\alpha_2^e\left(1-\alpha_6^f\right)B_1$

(17,9)	{1,2,1}	{0,2,2}	$\mu_2\alpha_2^e\alpha_6^f B_1$
(17,14)	{1,2,1}	{1,1,1}	$\mu_2\left(1-\alpha_6^f\right)(1-B_1)$
(17,15)	{1,2,1}	{1,1,2}	$\mu_2\alpha_6^f(1-B_1)$
(17,16)	{1,2,1}	{1,2,0}	$\mu_3\alpha_6^e$
(17,18)	{1,2,1}	{1,2,2}	$\mu_2\alpha_6^f(1-\alpha_2^e)$
(17,26)	{1,2,1}	{2,2,1}	$\gamma_1\ \alpha_2^f$
(18,9)	{1,2,2}	{0,2,2}	$\mu_3 B_3\alpha_3^e$
(18,15)	{1,2,2}	{1,1,2}	$\mu_3 B_4$
(18,17)	{1,2,2}	{1,2,1}	$\mu_3(1-B_1)$
(18,27)	{1,2,2}	{2,2,2}	$\gamma_1\ \alpha_3^f$
(19,13)	{2,0,0}	{1,1,0}	μ_1
(19,20)	{2,0,0}	{2,0,1}	γ_3
(19,22)	{2,0,0}	{2,1,0}	γ_2
(20,14)	{2,0,1}	{1,1,1}	μ_1
(20,19)	{2,0,1}	{2,0,0}	$\mu_3\alpha_6^e$
(20,21)	{2,0,1}	{2,0,2}	$\gamma_3\alpha_6^f$
(20,23)	{2,0,1}	{2,1,1}	γ_2
(21,15)	{2,0,2}	{1,1,2}	μ_1
(21,20)	{2,0,2}	{2,0,1}	μ_3
(21, 24)	{2,0,2}	{2,1,2}	γ_2
(22,13)	{2,1,0}	{1,1,1}	$\mu_1(1-\alpha_4^f)$
(22,16)	{2,1,0}	{1,2,0}	$\mu_1\alpha_4^f$
(22,20)	{2,1,0}	{2,0,1}	$\mu_1\alpha_4^e$
(22,23)	{2,1,0}	{2,1,1}	$\mu_1(1-\alpha_4^e)$
(22,25)	{2,1,0}	{2,2,0}	$\gamma_2\alpha_4^f$
(23,14)	{2,1,1}	{1,1,1}	$\mu_1(1-\alpha_4^f)$
(23,17)	{2,1,1}	{1,2,1}	$\mu_1\alpha_4^f$
(23,20)	{2,1,1}	{2,0,1}	$\mu_2(1-\alpha_6^f)\alpha_4^e$
(23,21)	{2,1,1}	{2,0,2}	$\mu_2\alpha_4^e\alpha_6^f$
(23,22)	{2,1,1}	{2,1,0}	$\mu_3\alpha_6^e$
(23,24)	{2,1,1}	{2,1,2}	$\mu_2(1-\alpha_4^e)\alpha_6^f$
(23,26)	{2,1,1}	{2,2,1}	$\gamma_2\alpha_4^f$
(24,15)	{2,1,2}	{1,1,2}	$\mu_1(1-\alpha_5^f)$
(24,18)	{2,1,2}	{1,2,2}	$\mu_1\alpha_5^f$
(24,21)	{2,1,2}	{2,0,2}	$\mu_3 B_2\alpha_5^e$
(24,23)	{2,1,2}	{2,1,1}	$\mu_3(1-B_2)$
(24,27)	{2,1,2}	{2,2,2}	$\gamma_2\alpha_5^f$
(25,19)	{2,2,0}	{2,0,0}	$\mu_2 B_1$
(25,23)	{2,2,0}	{2,1,1}	$\mu_2(1-B_1)$

(25,26)	{2,2,0}	{2,2,1}	γ_3
(26,17)	{2,2,1}	{1,2,1}	$\mu_2 B_1\,(1-\alpha_6^f)$
(26,18)	{2,2,1}	{1,2,2}	$\mu_2 B_1\,\alpha_6^f$
(26,23)	{2,2,1}	{2,1,1}	$\mu_2(1-B_1)\,(1-\alpha_6^f)$
(26,24)	{2,2,1}	{2,1,2}	$\mu_2(1-B_1)\,\alpha_6^f$
(26,25)	{2,2,1}	{2,2,0}	$\mu_3\alpha_6^e$
(26,27)	{2,2,1}	{2,2,2}	$\gamma_3\alpha_6^f$
(27,18)	{2,2,2}	{1,2,2}	$\mu_3 B_3$
(27,24)	{2,2,2}	{2,1,2}	$\mu_3 B_4$
(27,26)	{2,2,2}	{2,2,1}	$\mu_3(1-B_1)$

The diagonal elements where not includes in the table above. The value of the diagonal element is the negative sum of all elements in that row except the diagonal.

Bibliography

Abou-El-Ata, M., Al-seedy, R., Kotb, K. (1993). A transient solution of the state-dependent queue: M/M/1/N with balking and reflecting barrier. *Microelectronics Reliability,* 33(15): 681-688.

Cao, W., Stewart, W. (1985). Iterative aggregation/disaggregation techniques for nearly uncoupled Markov chain. *Journal of the Association of Computing Machinery,* 32(3): 702-719.

Clarke, A. (1956). A waiting line process of Markov type. *The Annals of Mathematical Statistics,* 27(2): 452-459.

Filipiak, J. (1988). Modelling and control of dynamic flows in communication networks. *Springer-Verlag, ISBN 3-540-18292-6*.

Grassman, W. (1977). Transient solutions in Markovian queueing systems. *Computers & Operations Research*, 4(1): 47-56.

Hogg, R., Tanis, E. (2006). Probability and statistical inference, volume 7e. *Pearson Education, Inc., Upper Saddle River*.

Kaczynski, W., Leemis, L., and Drew, J. (2012). Transient queueing analysis. *INFORMS Journal on Computing,* 24(1): 10-28.

Karpelevitch, F., Kreinin, A. (1992). Joint distributions in Poissonian tandem queues. *Queueing Systems,* 12(3-4):273-286.

Keilson, J. (1966). The ergodic queue length distribution for queueing systems with finite capacity. *Journal of the Royal Statistical Society, Series B*, 28(1):190-201.

Kolssar, P., Rider, K., Crabill, T., Walker, W. (1975). A queueing-linear programming approach to scheduling police patrol cars. *Operations Research*, 23(6):1045-1062.

Koopman, B. (1972). Air-terminal queues under time-dependent conditions. *Operations Research,* 20(6): 1089-1114.

Leese, E., Boyd, D. (1966). Numerical methods of determining the transient behavior of queues with variable arrival rates. *Canadian Operational Research Society Journal*, 4(1): 1-13.

Little, J. (1961). A proof for the queueing formula $L = \lambda W$. *Operations Research,* 9(3): 383-387.

Little, J. (2011). Little's law as viewed on its 50th anniversary. *Operations Research,* 59(3): 383-387.

Luchak, G. (1956). The solution of the single-channel queueing equations characterized by a time dependent arrival rate and a general class of holding times. *Operations Research,* 4(6): 711-732.

Morse, P. (1958). Queues, inventories and maintenance. The analysis of operational systems with variable demand and supply. *Wiley, New York*, pp. 65-67.

Muppala, J., Trivedi, K. (1992). Numerical transient solution of finite Markovian queueing systems. *Queueing and Related Models, U.N Bhat I. V. Basawa (ed.), Oxford University Press*, pp. 262-284.

Neuts, M. (1973). The single server queue in discrete time analysis I. *Naval Research Logistics Quarterly,* 20(2): 297-304.

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

Osorio, C. and M. Bierlaire. (2009b). A surrogate model for traffic optimization of congested networks: an analytic queueing network approach. *Ecole Polytechnique Federale de Lausanne*. Report TRANSP-OR 090825

Osorio, C. (2010) Mitigating network congestion: analytical models, optimizations methods and their applications", Doctoral Thesis, *Ecole Polytechnique Federale de Lausanne*.

Osorio, C., Flotterod, G., and Bierlaire, M. (2011). A differentiable dynamic network loading model that yields queue length distribution and account for spillback. *Transportation Research Part B*.

Osorio, C., Flotterod, G. (2012). Capturing dependency among link boundaries in stochastic dynamic network loading model. *Proceedings of the International Symposium on Dynamic Traffic Assignment (DTA).*

Osorio, C., Wang, C. (2012). An analytical approximation of the joint distribution of aggregate queue-lengths in urban network. *Proceedings of the Euro Working Group on Transportation Meeting*.

Parthasarathy, P. (1987). A transient solution to a M/M/1 Queue – A simple Approach. *Advanced applied probability*, 19(1): 997,998.

Phillips, G. (1995). Transient models for queueing networks. Masters thesis, *University of Stellenbosch*.

Rothkopf, M., Oren, S. (1979). A closure approximation for the nonstationary M/M/s queue. *Management Science,* 25(6): 522-534.

Schweitzer, P. (1984). Aggregation methods for large Markov chains. *Proceedings of the International Workshop on Computer Performance and Reliability,* p.275-285, North-Holland, Amsterdam.

Schweitzer, P. (1991). A survey of aggregation-disaggregation in large Markov chains. *Numerical Solution of Markov Chains, p.* 63-88, New York.

Stern, T. (1979). Approximations of queue dynamics and their applications to adaptive routing in computer communication networks. *IEEE Transactions on Communication,* 27(9):1331- 1335.

Strugul, John R. (2000). Mine design: examples using simulation. ISBN 0-87335-191-9

Takacs, Lajos. (1961). The transient behavior of a single server queueing process with a Poisson input. P. 535-567, *University of California Press*.

Takahashi, Y. (1975). A lumping method for numerical calculations of stationary distributions of Markov chains. *Research Reports on Information Sciences, Series B: Operations Research*.

Takahashi, Y. (1985). A new type aggregation method for large Markov chains and its application to queueing networks. In *proceedings of the International Teletraffic Congress 11*, Kyoto, Japan

Takahashi, Y., Song, Y. (1991). Aggregate approximation for tandem queueing systems with production blocking. *Journal of the Operations Research Society of Japan,* 34(3): 329-353.

TSS (2011). AIMSUN 6.1 *Microsimulator Users Manual.* Transport Simulation Systems.

VSS (1992). *Norme Suisse SN 640837 Installations de feux de circulation; temps transitoires et temps minimaux.* Union des professionnels suisses dela route, VSS, Zurich.

Wragg, A. (1963). The solution of the infinite set of differential- difference equations occurring in polymerization and queueing problems. *Mathematical Proceedings of the Cambridge Philosophical Society,* 59(1): 117-124.