Analysis of 802.11 DCF with Heterogeneous Non-Saturated Nodes

Hamed M. K. Alazemi
Dept. of Computer Engineering
Kuwait University
Kuwait
hamed@eng.kuniv.kw

A. Margolis, J. Choi,* R. Vijayakumar, S. Roy †
Dept. of Electrical Engineering
U. Washington Box 352500
Seattle, WA 98195

B. D. Choi
Telecommunication Mathematics Research Center and Department of Mathematics
Korea University
1, Anam-dong, Sungbuk-ku, Seoul 136-701, Korea

UWEE Technical Report
Number UWEEETR-2006-0004
February 2006
Abstract

In this paper we present a queueing model for a 802.11 wireless LAN under non-saturated conditions. Our model builds on previous work, but additionally takes into account heterogeneous arrival rates. In particular, we consider the case when there are two packet arrival rates in the network, with one set of stations generating packets at the lower rate and the rest generating packets at the higher rate.

1 Introduction

The successful deployment of IEEE 802.11 Wireless LANs has generated much interest in analytical models of network performance. The various physical layer standards that comprise the 802.11 family of protocols all rely upon the same medium access control (MAC) protocol ([1]), which is based on CSMA/CA. The performance of the 802.11 MAC in the so-called “saturation” mode (i.e. all stations always have data to send) has been widely studied in the literature [2], [3], [8], [17].
Bianchi [2] developed a Markov chain model from which he obtained expressions for the saturation throughput. His work (as also the work of Cali et al. [3]) showed that, in saturation, the behavior of 802.11 is well approximated by CSMA Slotted Aloha ([5]) with an appropriately chosen value of the transmission probability $p$. Chatzimisios et al. [8] built upon Bianchi’s Markov chain model to compute delays in the saturation case. Carvalho et al. [17] derived the average delay based on Bianchi’s model and results. They used a probabilistic approach to analyze the average service time for each packet, then derived the one hop delay of the packet.

Since saturation may be viewed as the limiting mode of operation when the arrival rates at all nodes tends to infinity, the “non-saturation” mode may equivalently be considered as the “finite rate” or “finite load” mode of operation. One approach to studying the finite rate mode of operation is to extend Bianchi’s Markov model by introducing additional states to model the operation of the system when some of the clients have no data to transmit. This is the approach taken by Cantieni et al. [9] and Ergen and Varaiya [10]. A different approach has been proposed by Foh and Zukerman [11] who use a queueing model with state dependent service rates; these rates are obtained from Bianchi’s saturation analysis. This idea of using rates from the saturation analysis was also used by Winands et al. in [18] to analyze 802.11 using the machine-repair model from queueing theory.

The limitation of using rates from the saturation analysis was overcome in recent work by Tickoo et al. [12] who introduced a different approach to study the performance analysis in non-saturation state by introducing probability generating functions, which allow the computation of the probability distribution function (pdf) of the delay. A detailed analytic model based on a discrete time G/G/1 queue which allows for the evaluation of the networks under consideration for general traffic arrival patterns and arbitrary number of users is proposed. Li et al. [14, 15, 16] proposed a simple model to analyze the performance of IEEE 802.11 DCF with service differentiation support in non-saturation states. In [13], a three dimensional Markov Chain model is introduced to attempt to describe the behavior of multi-hop 802.11 networks. Their argument lies on the fact that in most of the the work that has been done on the performance analysis of the IEEE 802.11 protocol, all stations behave like saturated data sources. For multi-hop wireless networks that include both source stations and relay stations, the average queue length, average delay, and energy consumption at relay stations are of great interest. Accordingly, to mathematically analyze those performance features, they introduce a third dimension in the chain that takes queue length into account.

The non-saturation analysis in [18, 12, 13, 14, 15, 16] deals with the symmetric case in which all clients have the same arrival rate. Our work is distinguished by the fact that we allow for asymmetric finite rate clients; i.e. we study the operation of the system in the non-saturation case when different clients have different data rates. This is clearly a model of considerable practical interest since, in real world systems, different clients could have widely varying traffic rates. In particular, most 802.11 deployments use the infrastructure mode of operation in which all traffic is routed through an access point; this access point will typically have much more traffic to transmit than the clients, but is considered a peer of the clients in the operation of the MAC protocol. In particular, most 802.11 deployments use the infrastructure mode of operation in which all traffic
is routed through an access point; this implies that access points will have much more traffic to transmit than the clients. However, in the current DCF protocol (infrastructure mode IBSS operation of 802.11 network), the AP is not accorded any priority but is considered a peer of the clients for purposes of channel access.

The paper is organized as follows. Section 2 develops the queueing model for DCF under non-saturation conditions; the analytical results are contained in Sections 3 and 4. Some numerical results are presented in Section 5.

2 Model Formulation

Our model is an extension of that described in [18], which is an adaptation of the basic homogeneous finite-source machine repair model that has been studied extensively in the literature [6], Kleinrock [4], and Jaiswal [7]. The machine repair problem with heterogeneous types has been discussed by [22], [19], [23] and [20, 21].

In this paper we assume that there are $N_1(N_2)$ stations of type 1(2) and one server. At any time a station can be either active or inactive. Inactive stations will generate a message after an exponentially distributed duration, with rate $\lambda_1(\lambda_2)$ for Type 1(2) stations, independently. Each message consists of a random, geometrically distributed number of packets with mean $1/q$. The time needed to transmit a single packet, called the service time, is exponentially distributed with rate $\mu$. A station that is currently transmitting a packet is said to be in service; a station that is active but not transmitting is said to be in the waiting room. Note that at any point in time, if there are $i(j)$ active type 1(2) stations, there must be $N_1 - i(N_2 - j)$ inactive type 1(2) stations, $i + j - 1$ stations in the waiting room, and one station in service (either type 1 or type 2). We assume that the service times are mutually independent and do not depend on the station type and the number of currently active stations. The order of service is random: after a service completion, the next station to begin service is equally likely to be any of the stations in the waiting room. When a station completes service it either returns to the waiting room with probability $q$ (it has another packet), or becomes inactive with probability $1 - q$. Note that the probability of a station returning to the waiting room is independent of how long that station has been active (or in service). We interpret this to mean that a station does not generate any new messages while it is active. To the best of our knowledge, this model has not been considered before in the open literature.

Remark: In the machine repair model of [18] seen here in Figure 1, machines break down independently with rate $\lambda_1$ and are serviced in random order with service rate $\mu$; a machine will be successfully fixed during service with probability $(1 - q)$, or will remain broken and return to the waiting room with probability $q$. Our model has a similar interpretation, except that now there are two types of machines which break down independently at different rates, $\lambda_1$ and $\lambda_2$. Their service rates, however, are identical.
3 Steady-state distribution

Let $X(t), Y(t)$ be random variables representing the number of active stations of type 1 and type 2 respectively at time $t$. In order to analyze the system as a Markov process, we define an additional random variable $K(t)$ to indicate the type of the station currently in service, in case there are some active stations of both types. The process $\{X(t), Y(t), K(t)\}$ then forms a three-dimensional continuous-time Markov chain with state space $S = \{(i, j, k) \mid 0 \leq i \leq N_1, 0 \leq j \leq N_2, k \in \{1, 2\}\}$. Note that the states $(i,0,2)$ and $(0,j,1)$ cannot occur. If no station is currently in service (which happens only when no stations are active), the value of $k$ is does not matter; we call this state $(0,0)$. For simplicity in what follows, $(0,0,1), (0,0,2),$ and $(0,0,0)$ are used as equivalent notation for state $(0,0)$.

Let $q(i_1,j_1,k_1),(i_2,j_2,k_2)$ denote the rate at which the system in state $(i_1,j_1,k_1)$ transitions into state $(i_2,j_2,k_2)$. Under the model assumptions, these one-step transition rates are:

$$q(1,0,1),(0,0,0) = q(0,1,2),(0,0) = \mu(1 - q)$$

$$q(i_1,j_1,0),(i_1-1,0,2) = \mu(1 - q) \quad \forall i, j \geq 1$$

$$q(i,j,X),(i+1,j,X) = (N_1 - i)\lambda_1 \quad \forall 0 \leq i \leq N_1 - 1$$

$$q(i,j,X),(i,j+1,X) = (N_2 - j)\lambda_2 \quad \forall 0 \leq j \leq N_2 - 1$$

$$q(i,j,1),(i,j,2) = \frac{i}{i+j} \mu q \quad \forall i, j \geq 1$$

$$q(i,j,2),(i,j,1) = \frac{j}{i+j} \mu q \quad \forall i, j \geq 1$$

$$q(i,j,1),(i-1,j,1) = \frac{i-1}{i+j-1} \mu(1 - q) \quad \forall i, j \geq 1$$

$$q(i,j,1),(i-1,j,2) = \frac{j}{i+j-1} \mu(1 - q) \quad \forall i, j \geq 1$$

$$q(i,j,2),(i,j-1,1) = \frac{i}{i+j-1} \mu(1 - q) \quad \forall i, j \geq 1$$

$$q(i,j,2),(i,j-1,2) = \frac{j-1}{i+j-1} \mu(1 - q) \quad \forall i, j \geq 1$$
Figure 2: Markov Chain Model
where \( X \) is either 1 or 2, and all other transitions cannot occur. The first two equations follow from the fact that the time for a station to complete transmission of one packet is exponential with rate \( \mu \), and with probability \((1 - q)\) that station has no more packets and turns off. The second two equations follow because if there are \( i \) type 1 and \( j \) type 2 stations active, there are \( N_1 - i \) type 1 stations that are inactive and \( N_2 - j \) type 2 stations that are inactive; therefore the time until the next arrival of type 1 is exponential with rate \((N_1 - i)\lambda_1\), and the time until the next arrival of type 2 is exponential with rate \((N_2 - j)\lambda_2\). The last four equations follow because the next station selected to transmit after the current transmission is equally likely to be any of the remaining active stations. It should be emphasized that these transition rates explicitly assume a random service discipline; under any other discipline the process \((X(t), Y(t), K(t))\) would not constitute a Markov process.

Let \( \pi_{(i,j,k)} = \lim_{t \to \infty} P\{(X(t), Y(t), K(t)) = (i, j, k)\} \) be the stationary distribution of this Markov chain. Since the state space is finite and all states communicate, these limiting probabilities exist and satisfy the set of balance equations given in appendix A. It can be shown using these balance equations and the normalization condition \(\sum_{(i,j,k)} \pi_{(i,j,k)} = 1\) that the stationary probabilities for this Markov chain are given by:

\[
\pi_{(0,0)} = \left(1 + \sum_{(i,j,k) \neq (0,0)} \left(\frac{\lambda_1}{\mu(1-q)}\right)^i \left(\frac{\lambda_2}{\mu(1-q)}\right)^j \left(\frac{N_1}{i}\right) \left(\frac{N_2}{j}\right) (i + j)! \Gamma(i,j,k)\right)^{-1}
\]

\[
\pi_{(i,j,k)} = \pi_{(0,0)} \left(\frac{\lambda_1}{\mu(1-q)}\right)^i \left(\frac{\lambda_2}{\mu(1-q)}\right)^j \left(\frac{N_1}{i}\right) \left(\frac{N_2}{j}\right) (i + j)! \Gamma(i,j,k), \forall (i + j) \geq 1
\]

where \(\Gamma(i,j,k) = \{i, \text{if } k=1\} \{j, \text{if } k=2\}\)

By summing over the type \((k)\) of the station in service, we get an expression for \(\pi_{(i,j)} = \lim_{t \to \infty} P\{(X(t), Y(t)) = (i, j)\}\), the distribution of the number of active stations of each type:

\[
\pi_{(i,j)} = \pi_{(i,j,1)} + \pi_{(i,j,2)} = \pi_{(0,0)} \left(\frac{\lambda_1}{\mu(1-q)}\right)^i \left(\frac{\lambda_2}{\mu(1-q)}\right)^j \left(\frac{N_1}{i}\right) \left(\frac{N_2}{j}\right) (i + j)!
\]

Let \(N(t) = X(t) + Y(t)\) be the total number of active stations at time \(t\). Then \(\pi_n = \lim_{t \to \infty} P\{N(t) = n\}\) is given by:

\[
\pi_n = \sum_{l=0}^{n} \pi_{(n-l)} = \frac{n! \pi_{(0,0)}}{(\mu(1-q))^n} \sum_{l=0}^{n} \lambda_1^l \lambda_2^{n-l} \left(\frac{N_1}{l}\right) \left(\frac{N_2}{n-l}\right)
\]

As a side note, it can be shown that if \(\lambda_1 = \lambda_2 = \lambda\), the above reduces to the expression given in [18], with \(N_1 + N_2\) stations:

\[
\pi_n = \pi_0 \left(\frac{\lambda}{\mu(1-q)}\right)^n \frac{(N_1 + N_2)!}{(N_1 + N_2 - n)!}
\]
where $\pi_0 = \left( \sum_{l=0}^{N_1+N_2} \left( \frac{\lambda}{\mu (1-q)} \right)^l (N_1+N_2)! \right)^{-1}$

### 4 Performance measures

We can easily derive the long-run marginal distributions for the number of active stations of each type:

$$\pi^x_i \equiv \lim_{t \to \infty} P\{X(t) = i\} = \pi_{(0,0)} \left( \frac{\lambda_1}{\mu (1-q)} \right)^i \binom{N_1}{i} \Theta_2(i)$$

$$\pi^y_j \equiv \lim_{t \to \infty} P\{Y(t) = j\} = \pi_{(0,0)} \left( \frac{\lambda_2}{\mu (1-q)} \right)^j \binom{N_2}{j} \Theta_1(j)$$

where $\Theta_1(j) \equiv \sum_{i=0}^{N_1} \left( \frac{\lambda_1}{\mu (1-q)} \right)^i \binom{N_1}{i} (i + j)!$

and $\Theta_2(i) \equiv \sum_{j=0}^{N_2} \left( \frac{\lambda_2}{\mu (1-q)} \right)^j \binom{N_2}{j} (i + j)!$

Now expressions for $E[\pi^x_i]$ and $E[\pi^y_j]$, the mean numbers of active stations of each type, can be derived from the above expressions according to:

$$E[\pi^x_i] = \pi_{(0,0)} \sum_{i=0}^{N_1} \left( \frac{\lambda_1}{\mu (1-q)} \right)^i \binom{N_1}{i} \Theta_2(i)$$

$$E[\pi^y_j] = \pi_{(0,0)} \sum_{j=0}^{N_2} \left( \frac{\lambda_2}{\mu (1-q)} \right)^j \binom{N_2}{j} \Theta_1(j)$$

Since $N_1 - E[\pi^x_i]$ and $N_2 - E[\pi^y_j]$ are the long-run average number of inactive type 1 and type 2 stations respectively, the throughput, or average rate at which stations become active, is given by:

$$\gamma = \lambda_1 (N_1 - E[\pi^x_i]) + \lambda_2 (N_2 - E[\pi^y_j])$$

Now $E[D]$ is just the average wait time for a station to send a message, that is, it is the average length of a station active period. According to Little’s Law, the average number of active stations is equal to the average activation rate times the average length of a station active period, so:

$$E[D] = \frac{E[\pi^x_i] + E[\pi^y_j]}{\gamma}$$

Let $D_1$ and $D_2$ be the random delays for a message from each type of station. Then the average delays for type 1 and type 2 messages are given by:

$$E[D_1] = \frac{E[\pi^x_i]}{\gamma_1}$$
\[ E[D_2] = \frac{E[\pi_y^y]}{\gamma_2} \]

where \( \gamma_1 = \lambda_1(N_1 - E[\pi_x^x]) \) and \( \gamma_2 = \lambda_2(N_2 - E[\pi_y^y]) \).

We next investigate the distribution of idle and busy periods. Note that the sequence of idle
and busy periods is an alternating renewal process, and satisfies:

\[ \pi_{(0,0)} = \frac{E[I]}{E[B] + E[I]} \]

where \( I \) is a random variable representing the length of an idle period and \( B \) the length of a busy
period. At the beginning of the idle period, there are \( N_1 \) inactive type 1 stations and \( N_2 \) inactive

\[ \text{type 2 stations, so the duration until the next busy period is just the time until the next activation,} \]

which is exponentially distributed with rate \( N_1 \lambda_1 + N_2 \lambda_2 \). Therefore \( E[I] = \frac{1}{N_1 \lambda_1 + N_2 \lambda_2} \), and we

\[ E[B] = (1 - \pi_{(0,0)}) \frac{1}{(N_1 \lambda_1 + N_2 \lambda_2) \pi_{(0,0)}} \]

We can immediately make a few observations about system performance based on these derivations. First, for fixed \( N_1, N_2, \lambda_1, \) and \( \lambda_2 \), mean delay and busy period both decrease when \( \mu(1-q) \) increases, while throughput increases. Second, as either \( \lambda_1 \) or \( \lambda_2 \) approaches infinity, the mean busy period approaches infinity. Third, the throughput is bounded by \( \mu(1-q) \), and approaches this limit in the limit of large arrival rates. This is readily understood since at high arrival rates, the system is almost always busy, and the overall rate at which stations become inactive will approach the rate during a busy period, \( \mu(1-q) \). Finally, the mean delay is bounded by \( (N_1 + N_2)/(\mu(1-q)) \), since the highest delay occurs then both arrival rates are infinite, and the rate at messages are sent when all stations are active is \( (\mu(1-q))/(N_1 + N_2) \).

5 Numerical Results

We consider how performance is affected by the relative arrival rates \( \lambda_1, \lambda_2 \) and relative numbers of
stations of each type, for a fixed network size \( N_1 + N_2 \) and traffic intensity \( \Lambda = N_1 \lambda_1 + N_2 \lambda_2 \). The
parameters used in this section were selected based on those in [18]. Two different network sizes
are considered, under two different values of the system load \( \Lambda/(\mu(1-q)) \). Figures 3 and 4 show the
mean delay \( E[D] \), throughput, and mean busy period \( E[B] \) vs. the log of the ratio \( R = \lambda_1/\lambda_2 \), for
network sizes 10 and 25, and for loads 0.25 and 8.0. The service rates are \( \mu = 1/197.6 \) for the size
10 network and \( \mu = 1/196.4 \) for the size 25 network; in both cases the average number of packets in
a message is 20. Figures 5 and 6 show the mean delays for type 1 and 2 stations, \( E[D_1] \) and \( E[D_2] \),
under the same conditions. For each plot, curves are shown for networks composed of different
numbers \( N_1, N_2 \) of each station type. Note that for a fixed load and network size, \( R = 1 \) represents
a network composed of \( N_1 + N_2 \) identical stations, each with activation rate \( \lambda = \Lambda/(N_1 + N_2) \); hence all curves pass through the same point at \( R = 1 \).
From these plots, we observe that in all cases, $E[D]$, throughput, and $E[B]$ tend to be larger when the stations have similar activation rates. This is reasonable since as $R$ goes to 0, the network approaches a network with only $N_2$ stations (each with rate $\lambda/N_2$), and as $R$ goes to infinity, the network approaches a network with only $N_1$ stations (each with rate $\lambda/N_1$). In the case with the highest load and network size (8.0, 25), the throughput has essentially reached its bound of $\mu(1-q)$ for all network compositions, and does not noticeably change until $\lambda_1$ and $\lambda_2$ are very different. However, $E[D]$ is still significantly below its bound of $(N_1 + N_2)/(\mu(1-q))$, even at $R = 1$. Under a smaller network size with the same load, the throughput is much more sensitive to changes in $R$. We also observe that $E[B]$ is generally more sensitive to changes in $R$ at high values of the load, especially in the larger network. We note the interesting result that $E[D]$, throughput, and $E[B]$ appear to have the same dependence on $R$ for the small load cases.

For a fixed network size, the effect of network composition depends on the load. Under a high load, $E[D]$ and $E[B]$ for networks composed mostly of one type of station (for example, 8/2 and 24/1) appear to be less sensitive to changes in $R$ than the more heterogeneous networks. In particular, changes in $R$ have very little effect on either the individual delays $E[D_1], E[D_2]$ or the overall delay $E[D]$ under the high load 24/1 network, and $E[B]$ also shows less sensitivity to $R$ than in the other three networks. This is not true for the small load networks. For $\lambda_2 > \lambda_1$ under a small load, the most heterogeneous networks (5/5, 13/12) are in fact the least sensitive to changes in small $R$. We can expect that decreasing $R$ will initially have little effect on a network composed mostly of type 1 stations, but at some point, depending on the load, the decreasing value of $\lambda_1$ will result in the system becoming much less busy than the other networks, due to the larger number of type 1 stations. This accounts for the sudden drop in the values of $E[D]$, throughput, and $E[B]$ in the 24/1 network compared with other networks under small load. This effect is not observable in the high load case, due to the fact that $\lambda_1$ is still quite large for this range of $R$, but we can expect to see it at a smaller value of $R$. All networks with more type 1 than type 2 stations are more sensitive to changes in $R$ when $\lambda_1 < \lambda_2$.

The plots of $E[D_1]$ and $E[D_2]$ show that, under small load, decreasing $R$ generally results in an increase of type 1 mean delay and a decrease in type 2 mean delay; but for high load, there is generally a decrease in both types of delay in both directions of $R$. Increasing $R$ can cause a decrease in delay for both types of stations at high load, since there are fewer active type 2 stations. However, the increased activation rate of type 1 stations can also result in an increase of $E[D_2]$ if there are few type 2 stations and a high load, as is apparent in the 24/1 case. Thus the effect depends on the load and network composition.
Figure 3: Mean delay, throughput, and mean busy period as a function of $\log(\frac{\lambda_1}{\lambda_2})$ for fixed network size $N_1 + N_2 = 10$ and fixed loads $\frac{\lambda_1}{\mu(1-\gamma)} = 0.25, 8.0$. 
Figure 4: Mean delay, throughput, and mean busy period as a function of $\log(\frac{\lambda_1}{\lambda_2})$ for fixed network size $N_1 + N_2 = 25$ and fixed loads $\frac{\Lambda}{\mu(1-\eta)} = 0.25, 8.0$. 

11
Figure 5: Mean delay for each station type as a function of $\frac{\lambda_1}{\lambda_2}$ for fixed network size $N_1 + N_2 = 10$ and fixed loads $\frac{\Lambda}{\mu(1-q)} = 0.25, 8.0$. 

\[ \]
Figure 6: Mean delay for each station type as a function of $\frac{\lambda_1}{\lambda_2}$ for fixed network size $N_1 + N_2 = 25$ and fixed loads $\frac{\Lambda}{\mu(1-q)} = 0.25, 8.0$. 
Figure 7: Expected number of active type 1 and type 2 stations as a function of $\frac{\lambda_1}{\lambda_2}$ for fixed network size $N_1 + N_2 = 10$ and fixed loads $\frac{\lambda}{\mu(1-q)} = 0.25, 8.0$. 
Figure 8: Expected number of active type 1 and type 2 stations as a function of $\frac{\lambda_1}{\lambda_2}$ for fixed network size $N_1 + N_2 = 25$ and fixed loads $\frac{\Lambda}{\mu(1-q)} = 0.25, 8.0$. 
References


Appendix A: Balance Equations

\[(N_1 \lambda_1 + N_2 \lambda_2) \pi_{(0,0,0)} = (\mu (1 - q)) \pi_{(0,1,0)} + (\mu (1 - q)) \pi_{(0,1,2)} \]

\[(N_2 \lambda_2 + \mu (1 - q)) \pi_{(N_1,0,1)} = (\mu (1 - q)) \pi_{(N_1,1,2)} + (\lambda_1) \pi_{(N_1-1,0,1)} \]

\[(N_1 \lambda_1 + \mu (1 - q)) \pi_{(0,N_2,2)} = (\mu (1 - q)) \pi_{(1,N_2,1)} + (\lambda_2) \pi_{(0,N_2-1,2)} \]

\[
\begin{align*}
((N_1 - i) \lambda_1 + N_2 \lambda_2 + \mu (1 - q)) \pi_{(i,0,1)} &= (\mu (1 - q)) \pi_{(i+1,0,1)} + (\mu (1 - q)) \pi_{(i,1,2)} \\
&\quad + ((N_1 - i + 1) \lambda_1) \pi_{(i-1,0,1)} \\
\forall 0 < i < N_1
\end{align*}
\]

\[
\begin{align*}
\left( (N_2 - j) \lambda_2 + \mu (1 - q) + \frac{j}{N_1 + j} \mu q \right) \pi_{(N_1,j,1)} &= \left( \mu (1 - q) \frac{N_1}{N_1 + j} \right) \pi_{(N_1,j+1,2)} + (\lambda_1) \pi_{(N_1-1,j,1)} \\
&\quad + ((N_2 - j - 1) \lambda_2) \pi_{(N_1,j-1,1)} + \left( \mu q \frac{N_1}{N_1 + j} \right) \pi_{(N_1,j,2)} \\
\left( (N_2 - j) \lambda_2 + \mu (1 - q) + \frac{N_1}{N_1 + j} \mu q \right) \pi_{(N_1,j,2)} &= \left( \mu (1 - q) \frac{j}{N_1 + j} \right) \pi_{(N_1,j+1,2)} + (\lambda_1) \pi_{(N_1-1,j,2)} \\
&\quad + ((N_2 - j - 1) \lambda_2) \pi_{(N_1,j-1,2)} + \left( \mu q \frac{j}{N_1 + j} \right) \pi_{(N_1,j,1)} \\
\forall 0 < j < N_2
\end{align*}
\]

\[
\begin{align*}
(N_1 \lambda_1 + (N_2 - j) \lambda_2 + \mu (1 - q)) \pi_{(0,j,2)} &= (\mu (1 - q)) \pi_{(0,j+1,2)} + (\mu (1 - q)) \pi_{(1,j,1)} \\
&\quad + ((N_2 - j + 1) \lambda_2) \pi_{(0,j-1,2)} \\
\forall 0 < j < N_2
\end{align*}
\]

\[
\begin{align*}
\left( (N_1 - i) \lambda_1 + \mu (1 - q) + \frac{i}{N_2 + i} \mu q \right) \pi_{(i,N_2,2)} &= \left( \mu (1 - q) \frac{N_2}{i + N_2} \right) \pi_{(i+1,N_2,1)} + (\lambda_2) \pi_{(i,N_2-1,2)} \\
&\quad + \left( \mu q \frac{N_2}{i + N_2} \right) \pi_{(i,N_2,1)} \\
\left( (N_1 - i) \lambda_1 + \mu (1 - q) + \frac{N_2}{N_2 + i} \mu q \right) \pi_{(i,N_2,1)} &= \left( \mu (1 - q) \frac{i}{N_2 + i} \right) \pi_{(i+1,N_2,1)} + (\lambda_2) \pi_{(i,N_2-1,1)} \\
&\quad + \left( \mu q \frac{i}{N_2 + i} \right) \pi_{(i,N_2,2)}
\]

18
\[
\left(\frac{N_2}{N_1 + N_2} \mu q + \mu (1 - q)\right) \pi_{(N_1,N_2,1)} = \left(\frac{N_1}{N_1 + N_2} \mu q\right) \pi_{(N_1,N_2,2)} + ((N_1 - 1) \lambda_1) \pi_{(N_1-1,N_2,1)} + (N_2 - 1) \lambda_2 \pi_{(N_1,N_2-1,1)}
\]

\[
\left(\frac{N_1}{N_1 + N_2} \mu q + \mu (1 - q)\right) \pi_{(N_1,N_2,2)} = \left(\frac{N_2}{N_1 + N_2} \mu q\right) \pi_{(N_1,N_2,1)} + ((N_2 - 1) \lambda_2) \pi_{(N_1,N_2-1,2)} + (N_1 - 1) \lambda_1 \pi_{(N_1-1,N_2,2)}
\]

\[
\left(\frac{j}{i+j} \mu q + \mu (1 - q) + (N_1 - i) \lambda_1 + (N_2 - j) \lambda_2\right) \pi_{(i,j,1)} = \left(\frac{i}{i+j} \mu q\right) \pi_{(i,j,2)} + ((N_1 - i + 1) \lambda_1) \pi_{(i-1,j,1)} + ((N_2 - j + 1) \lambda_2) \pi_{(i,j-1,1)} + \left(\mu (1-q) \frac{i}{i+j}\right) \pi_{(i+1,j,1)} + \left(\mu (1-q) \frac{i}{i+j}\right) \pi_{(i,j+1,2)}
\]

\[
\forall 0 < i < N_1, 0 < j < N_2
\]

\[
\left(\frac{i}{i+j} \mu q + \mu (1 - q) + (N_1 - i) \lambda_1 + (N_2 - j) \lambda_2\right) \pi_{(i,j,2)} = \left(\frac{j}{i+j} \mu q\right) \pi_{(i,j,1)} + ((N_1 - i + 1) \lambda_1) \pi_{(i-1,j,2)} + ((N_2 - j + 1) \lambda_2) \pi_{(i,j-1,2)} + \left(\mu (1-q) \frac{j}{i+j}\right) \pi_{i+1,j,1} + \left(\mu (1-q) \frac{j}{i+j}\right) \pi_{i,j+1,2}
\]

\[
\forall 0 < i < N_1, 0 < j < N_2
\]