Performance Analysis of Synchronous Packet Networks with Priority Queueing Disciplines *

Audrey M. Viterbi
Department of Electrical Engineering
University of California
Irvine, CA 92717

Abstract
In this paper, models and analytical techniques are developed to evaluate the performance of time-synchronous packet networks with priority queueing disciplines. The theoretical results are validated by extensive computer simulation. The results of this research may be applied to the design of multiprocessor networks, local area networks, and cellular, terrestrial and satellite packet networks.

1. Introduction
A fundamental problem in the design of multi-user communication systems involves the merging, in a time-shared manner, of several users on a common medium, bus or network. Depending on the application field, this is referred to variously as multiplexing, multiple access, or demand assignment network allocation.

The inherently bursty nature of data communication, coupled with the need to share communication resources (primarily channel capacity) efficiently, suggests the packet-switching technique. When this form of switching is used, if a user has a block of data to be sent, it is divided into fixed-length packets which are stored at the first buffer and then forwarded later, one hop at a time. Each block is received in its entirety at each intermediate node and then retransmitted. The advantages of a fixed packet length are threefold: ease of storage of packets in transit because of prescribed buffer requirements; assurance of network fairness because unlike large messages, a fixed length packet cannot monopolize resources; and ability to forward the first packet of a multipacket message before subsequent packets have fully arrived, thereby reducing delay and improving throughput [1]. For these reasons, networks are usually packet-switched and slotted (where time is segmented into "slots" whose duration is equal to the duration of a single packet and all packets are required to begin their transmission at the beginning of a slot).

We consider priority in synchronous packet networks both for its own sake and as a form of flow control. We will treat a version of the flow control scheme termed Input Buffer Limit Flow Control. This scheme distinguishes between various kinds of input traffic and throttles input traffic based on entry node buffer occupancy. The most popular version of this scheme was proposed by Lam [3]. We consider two classes of traffic, input (exogenous) and transit traffic, and assign priority to transit traffic within selected networks.

Single-node priority systems, in which packets arrive according to a Poisson process and service time follows the exponential distribution, have been studied in the literature. More general service time distributions have also been considered [4,5,6,7]. Salomon [8] has considered a unidirectional tandem network without departures with deterministic service times. These results have been for continuous-time systems rather than discrete-time (or "slotted") systems.

In addition, little is known about arbitrary networks of priority queues. Packet lengths are often assumed to be exponentially distributed, so that Burke's theorem for an M/M/1 queueing system can be invoked to conclude that the packet departure process at the output

---

*This research was supported by NSF Grant NCR-8706282

CH2706-0/89/0000/0440$01.00 © 1989 IEEE
of a buffer is at steady-state a Poisson point process, as is the input. Then, to avoid statistical dependencies, an "independence assumption" is made by Kleinrock [9], which amounts to choosing the length of packets at random at each of the nodes in a network. For large and topologically complex continuous-time networks, these results are reasonable approximations. In our discrete-time situation, however, in the presence of simple topologies and constant length packets, this approximation is poor, and we need to perform a different analysis, as we shall show.

It is interesting to note that preemption is not an issue in the discrete-time priority systems described in this paper. In fact, discrete-time priority systems are nonpreemptive as no packets of higher priority will arrive during a packet's transmission. Discrete-time priority systems, however, share the property with continuous-time preemptive priority systems, that the packet in service is always of the highest priority present in the system.

In Section 2 we develop and analyze a local Markov model to approximate the delay incurred by each class of traffic in a network. In Section 3 we consider the application of this model to selected networks and show that it accurately predicts behavior of tandem (with and without departures), unidirectional and bidirectional loop networks. We are primarily interested in analyzing networks with simple, fixed topologies.

2. Mathematical Models for Time-Synchronous Packet Networks with Priority Queueing Disciplines

2.1 Poisson Model

Consider the buffers shown in Figure 1. User packets of different priority classes enter the appropriate buffer and await their turn to use the common medium. We adopt the convention that packets of priority class \( k \) have priority over packets of classes greater than \( k \).

We further assume that the arrivals of fixed-length packets into the priority buffers are due to a multiple access scheme and that they are independent bulk-Poisson processes (where \( j \) packets enter buffer \( i \) at the beginning of a slot, and the random variable \( j \) is Poisson distributed with mean arrival rate \( \alpha_i \)). More general input distributions will also be considered.

Since the unit of data is the packet, which is of fixed duration, the same constant amount of service will be required at each node. Thus we model the priority class buffers as infinite capacity queues with constant service times. We assume that the buffers of each priority class are governed by a first-come first-served (FCFS) policy. With the assumed slotted time structure, if buffers 1, 2, \( \ldots \), \( k-1 \) are empty and buffer \( k \) is nonempty, then one packet is removed from buffer \( k \) at equally spaced epochs \( t\Delta \), where \( \Delta \) is the time to service a packet.

We first consider the case of \( N \) independent priority classes feeding buffers 1, 2, \( \ldots \), \( N \) in bulk Poisson fashion at rates \( \alpha_i \), \( i = 1, \ldots, N \). If \( \sum_{i=1}^{k} \alpha_i < 1 (k = 1, \ldots, N) \) then we obtain the mean steady-state queueing delays of buffers 1, \( \ldots \), \( k \) as follows:

The mean steady-state queueing delay in buffer 1 is easily found [10] to be

\[
W_1 = \frac{\alpha_1}{2(1 - \alpha_1)}
\]

since class 1 behavior is independent of the arrivals of all other classes.

Further, since the priority queueing discipline does not affect the average packet delay, we can solve recursively for the mean steady-state queueing delay in buffer \( n \) [11,12]

\[
W_n = \frac{1}{2} \left[ \frac{1}{(1 - \sum_{i=1}^{n-1} \alpha_i)(1 - \sum_{i=1}^{n-1} \alpha_i)} - 1 \right],
\]

\( n = 1, \ldots, k \) (2)

We refer to this as the Poisson Model. A similar expression can be derived for general independent arrival processes.

2.2 Markov Model

We now consider an arbitrary network which is time-synchronous and operates in a store-and-forward manner. A packet arriving at an intermediate node is directed into the appropriate priority channel if that channel is free. If, instead, the channel is busy, the packet
joins a queue for access to that channel, which is governed by a first-come first-served (FCFS) policy. Thus, any contention that may occur manifests itself in the buffers and not on the links.

The performance of networks is typically measured in terms of throughput and delay. Since we assume that the buffers are of infinite capacity, which implies that there is no loss of packets, throughput is not degraded in the networks we study. Thus we are primarily concerned here with delay. Clearly, the extension to the study of priority classes within networks considerably complicates the problem. Now we must consider the correlation of delays between stages (or, equivalently, hops) in a network.

Central to our modeling of discrete-time distributed network is the development of a local model for the queueing delay of a each priority class at a given node. We then build our model of discrete-time, slotted networks (with equal packet lengths on all links) by linking several node buffers, each of which is fed by two or more packet streams whose characteristics depend on the behavior of the buffer from which they arrive. That is, we derive global network results from combinations of local node models.

We first consider applying the Poisson Model as our local model, but the inability of the Poisson Model to predict the behavior of arbitrary networks (as shown in the next section and in [11,12]) leads us to consider an alternative model. We develop a simple local model for queueing delay and find that it accurately predicts the behavior for a wide range of network topologies.

A very simplistic model would assume that the output of each queue is memoryless. That is, we might model each output sequence as a Bernoulli process with mean departure rate $a_i$. (In expanding to a network, the mean departure rates for the nodes are very easily computed by “conservation-of-flow” arguments on the network.) If we also assume the independence of the arrival process of each priority class, elementary queueing theory shows that the mean queueing delay of the $k$th priority class is

$$W_k = \left[ \frac{\sum_{i=1}^{k} a_i}{2(1 - \sum_{i=1}^{k} a_i)} \right] + \frac{\sum_{i=1}^{k-1} a_i}{2(1 - \sum_{i=1}^{k} a_i)}, k = 2, \ldots, N$$

(3)

The main weakness of this independent, Bernoulli network link model is that, since each link carries the output of a constant service-time queue, which at times may be filled with numerous packets, its output traffic tends to be quite bursty; that is, departures will tend to group together. Furthermore, since the nodes feed one another, arrivals may not be independent.

A model which provides a first-order approximation of such burstiness behavior is a first-order Markov model.

The main weakness of this independent, Bernoulli network link model is that, since each link carries the output of a constant service-time queue, which at times may be filled with numerous packets, its output traffic tends to be quite bursty; that is, departures will tend to group together. Furthermore, since the nodes feed one another, arrivals may not be independent.

Figure 2: First-Order Markov Chain

Such a model was successful in characterizing delays in a nonprioritized synchronous packet network [13,14]. We now model the dependence of buffer $k$ (L buffer) on buffers $i$, $1 \leq i \leq k - 1$ (H buffers) by the two-state Markov chain displayed in Figure 2, where

$$a_H = P[H \text{ buffers go from empty to nonempty in one time slot}]$$

$$b_H = P[H \text{ buffers remain nonempty in one time slot}]$$

$$a_H = E[\text{nonempty H buffers}]$$

$$= \sum_{i=1}^{k-1} a_i$$

We further define $\gamma_H = b_H - a_H$, which is the “burstiness parameter”, since for a Bernoulli sequence $\gamma_H = 0$. It is easily shown that, since the arrival rate $\alpha_H$ is related to $a_H$ and $b_H$ by

$$\alpha_H = \frac{a_H}{1 - \gamma_H}.$$  

(4)

We are primarily interested in how this Markov model predicts delay. In the appendix, we obtain the moment generating function of the delay distribution for the case of $N$ priority classes. From this, the mean steady-state queueing delay for the $k$th priority class is shown to be

$$W_k = \left[ \frac{\alpha_H + \alpha_L}{2(1 - \alpha_H - \alpha_L)} \right] + \frac{\alpha_H}{2(1 - \alpha_H - \alpha_L)} \left( 1 + \frac{1}{1 - \gamma_H} \right),$$

$$k = 2, \ldots, N$$

(5)
Note that the first bracketed term would be the delay for each packet if all users had the same priority in an isolated buffer. Hence prioritizing causes the kth class to have increased delay equal to a term involving sum of the arrival rates for all higher priority packets. This gives us an indication of the penalty to the kth class and also of the advantage of the higher priority classes (1, 2, ..., k - 1) as compared to a uniform service scheme without priorities. Clearly, when $\gamma_H = 0$, (6) reduces to the case where each priority class has packet departures which satisfy a Bernoulli distribution (3), independent of that of each other class.

Of great interest are the simulation results of Kruskal, Snir, and Weiss [15], which show that the correlation of delays between stages in a network drops off quickly for all but successive stages. This implies that our two-state model should perform well, given a good choice of the burstiness parameter, $\gamma$.

2.3 Choice of Burstiness Parameter for a Network

The choice of the burstiness parameter, $\gamma$, for an arbitrary network would appear to involve more than just the application of the results for an isolated node. The high priority buffers are fed by N streams of binary, though usually not Bernoulli, data. These streams are generally fractions of the outputs of buffers. Therefore we must deal with both splits of traffic at the output of a buffer and merges of traffic at the input to a buffer. This complicates the issue of quantifying the "burstiness" of the streams. This is particularly the case in an arbitrary network, where the traffic streams are the result of merging and splitting of non-Poisson traffic at the inputs and outputs of buffers at various stages (or equivalently hops) of the network. This changes the "burstiness" characteristic and hence possibly the relation of $\gamma_a$ to $a$. Nevertheless, we model the input streams to an individual buffer as independent and our simulation results show this assumption to be reasonably valid. In fact, in the case of tandem networks, and fixed routing, the input streams to an individual buffer are independent since they result from splitting and merging traffic originating from independent, exogenous sources.

We choose $\gamma$ as a function of $a$, because the $a$'s are calculated exactly by conservation-of-flow arguments. We show that a bulk-Poisson model for which,

$$\alpha_H \equiv 1 - e^{-\alpha_H}$$  \hspace{1cm} (6)
$$\gamma_H \equiv 1 - \frac{\alpha_H}{\alpha_H}$$  \hspace{1cm} (7)

appears to work well, except at very high rates (above rates of which are .9 of the link capacity). This is true even when we take into account merges and splits, which increase and reduce the burstiness, respectively, as shown in Figure 3. (As shown there, in the case of a binary split, a fraction $p$ of the traffic goes one way and a fraction $1-p$ goes the other way. Thus the expected burst length is reduced. In the case of the binary merge, while the input streams are not very bursty, the output stream may contain a steady stream of packets.) This choice of $a$ (and consequently $\gamma$) was used successfully in characterizing delays in nonprioritized synchronous packet networks [14,15]. The results of this choice of $\gamma$, along with simulation results, are presented in the next section for some simple, fixed topologies. In the following section, we test this heuristic for choosing the burstiness parameter on selected networks.

3. Application to Selected Networks

In this section, we consider a few simple, regular topologies which are often used for the architecture of a network. We compare simulation results with the approximation methods of Section 2.

3.1 Tandem Networks without Departures

Consider the unidirectional tandem network shown in Figure 4. Priority is given to transit traffic and the arrival processes are independent bulk-Poisson processes.

No queueing delay is suffered by the transit traffic, as at most one packet enters any of the transit buffers in one time slot. The "Poisson Model" of the previous section (3) is exact for the delay in the input buffer at each stage. This is obvious when all traffic is combined since the Poisson character of the transit traffic is retained.

3.2 Tandem Networks with Departures

We again consider the unidirectional tandem network shown in Figure 4, but we now allow departures at intermediate nodes. Priority is again given to transit traffic (thus no queueing delay is suffered by the transit traffic) and the arrival processes are independent bulk-Poisson
Table 1: Mean Queueing Delay in Input Buffers of a Tandem Network with Departures

<table>
<thead>
<tr>
<th>Node</th>
<th>$\alpha_H$</th>
<th>$\alpha_L$</th>
<th>Sim.</th>
<th>Markov</th>
<th>Poisson</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6</td>
<td></td>
<td>0.76</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
<td>0.4</td>
<td>3.59</td>
<td>3.43</td>
<td>3.67</td>
</tr>
<tr>
<td>3</td>
<td>0.4</td>
<td>0.4</td>
<td>3.33</td>
<td>3.43</td>
<td>3.67</td>
</tr>
</tbody>
</table>

Figure 4: Tandem Network without Departures

The simulation program was written in the C language. A different random number seed was used for each simulation. The simulations began with empty buffers, but the first 10,000 packets passing through a given buffer were not counted so as to minimise the effects of the initial transient. This choice was based on approximate guidelines recommended by Law [16]. The simulations were run for 1,000,000 packets. In all succeeding tables, in parentheses below the values given by the Markov and Bernoulli approximations, are the relative percent errors (R%E), which are defined as $R\%\mathrm{E} = 100\left(\frac{\text{Approximation} - \text{Simulation}}{\text{Simulation}}\right)$.

The mean queueing delay of node 1 is calculated exactly using (1). To distinguish this control data, it is designated by square brackets. The mean queueing delay of nodes 2 and 3 can be approximated using either the Markov or Poisson Model, but the Markov Model appears to be a better choice.

3.3 Unidirectional Loop Networks

Consider the unidirectional loop network shown in Figure 5. Again, priority is given to transit traffic and the arrival processes are independent bulk-Poisson processes. Now, however, transit traffic is dependent on the exogenous inputs.

To test the applicability of the independent Poisson model and Markov Model of the previous section to a unidirectional loop network, a N-node unidirectional loop network with uniform bulk-Poisson traffic $\alpha$ between all distinct source-destination pairs was simulated. Again, no queueing delay is suffered by the transit traffic in this configuration, as at most one packet enters any of the transit buffers in one time slot.

Table 2 contains the results of the simulation for $N=3,4,5,6$ and $10$. The high priority traffic (transit traffic), which is of rate $\frac{1}{2}(N-1)(N-2)\alpha$, is labeled H, while the low priority traffic (input traffic), which is of rate $(N-1)\alpha$, is labeled L. Table 2 compares the unidirectional loop simulation results of the previous section with the results calculated for the independent Poisson model of Section 2 and the Markov model (equation (6)) with Poisson arrival rate $\alpha_H = 1 - e^{-\alpha_H}$ and consequently $\gamma_H = 1 - (1 - e^{-\alpha_H})/\alpha_H$. The increasing discrepancies between the simulation results and the Poisson model with larger N (with consequently larger total high priority throughput) is due to bursty behavior which can not be modeled as a Poisson process. It is apparent that the
Markov model produces far more accurate results for all rates tested.

3.4 Bidirectional Loop Network

A bidirectional loop network is shown in Figure 6. Priority is again assigned to the transit traffic and the arrival processes are independent bulk-Poisson processes. The bidirectional loop is the only regular network we consider in this section that does have an obvious routing policy associated with it. We assume that a packet is routed along the minimum hop count path. In the case of equal length paths, the path is randomly chosen. Thus, if in a six node bidirectional loop we wish to send a packet from node 1 to node 4, with probability 1/2 we send it along route 1-2-3-4 and with probability 1/2 we send it along route 1-6-5-4.

The equation for the number of packets stored in the transit buffers in a bidirectional loop is similar to that of the unidirectional loop except that now internode traffic is split between two buffers which forward packets in opposite directions. Splits reduce burstiness as we saw in

Table 2: Mean Queueing Delay in Input Buffers of a Unidirectional Loop Network

<table>
<thead>
<tr>
<th>N</th>
<th>α</th>
<th>α_H</th>
<th>α_L</th>
<th>Mean Delay Sim. Markov Poisson</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.3</td>
<td>.3</td>
<td>.6</td>
<td>6.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.1%)</td>
</tr>
<tr>
<td>4</td>
<td>.15</td>
<td>.45</td>
<td>.45</td>
<td>7.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0%)</td>
</tr>
<tr>
<td>5</td>
<td>.09</td>
<td>.54</td>
<td>.36</td>
<td>8.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.3%)</td>
</tr>
<tr>
<td>6</td>
<td>.06</td>
<td>.6</td>
<td>.3</td>
<td>9.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.6%)</td>
</tr>
<tr>
<td>10</td>
<td>.02</td>
<td>.72</td>
<td>.18</td>
<td>11.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(7.4%)</td>
</tr>
</tbody>
</table>

Table 3: Mean Queueing Delay in Input Buffers of a Bidirectional Loop Network

<table>
<thead>
<tr>
<th>N</th>
<th>α</th>
<th>α_H</th>
<th>α_L</th>
<th>Mean Delay Sim. Markov Poisson</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>.18</td>
<td>.54</td>
<td>.36</td>
<td>8.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.3%)</td>
</tr>
<tr>
<td>6</td>
<td>.09</td>
<td>.45</td>
<td>.405</td>
<td>5.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.0%)</td>
</tr>
</tbody>
</table>

Figure 6, but we find that the first-order Markov model with \( \gamma = \gamma_{\text{Poisson}} \) is still a good approximation for the delays.

Table 3 contains the result of simulating an N-node bidirectional loop (for N=4 and 6) with uniform bulk-Poisson traffic of rate \( \alpha \rightarrow \gamma_{\text{Poisson}} \) for all unique source-destination pairs. The high priority traffic (transit traffic), which is of rate \( \frac{1}{2}N^2\alpha \) if N is even and \( \frac{1}{2}(N^2 - 1)\alpha \) if N is odd, is labeled H, while the low priority traffic (input traffic), which is of rate \( (N - 1)\alpha \), is labeled L. Again, no queueing delay is suffered by the transit traffic in this configuration, as at most one packet enters any of the transit buffers in one time slot.

It is again apparent that the Markov model produces far more accurate results. We note that it appears that the Markov Model approximation is a lower bound to the true delay as it is bounded above by the simulation results in all the topologies considered. Nevertheless, we believe that this Markov model is a good approximation when the rates (\( \alpha \)'s) of the arrival processes are not near capacity. When this is not the case, it is necessary to use a more sophisticated model involving more than two states.

We plan to test the applicability of this Markov Model to a wider range of network topologies. This will require the use of both the models developed in this paper (to treat the combining of traffic streams of different priorities) and those developed earlier for nonprioritized networks (to treat the combining of traffic streams of the same priority) [14,13].

4. Conclusion

In this paper, we have developed models and analytical techniques that were used to evaluate the performance of selected synchronous priority-oriented packet networks. In the cases which required approximation, we validated the model by performing extensive computer simulation and comparing with analytical results.

We developed a local first-order Markov model based on a "burstiness" criterion to approximate the delay incurred by each class of traffic in a network. Our simulations demonstrated the accuracy of the Markov model for a wide range of network topologies and arrival rates.
References


A. Appendix: Approximate First-Order Markov Model for Dependent Arrival Processes

Consider the buffers (queues) 1, . . . , N shown in Figure 1. If buffers 1, 2, . . . , k − 1 are empty and buffer k is nonempty, then one packet is removed from buffer k at rate αk at equally spaced epochs iΔ, where Δ is the time to service a packet. We define the interval [iΔ, (i + 1)Δ) as the ith slot. Then the number of packets stored in node buffer k at the end of slot i,

\[ S_k(i) = [S_k(i-1) - J]^+ + \alpha_k(i) \]  (A.1)

where \( \alpha_k(i) \) = number of packets arriving at buffer k during slot i

\[ I = \begin{cases} 1 & \text{if buffers } 1, \ldots, k-1 \text{ are empty} \\ 0 & \text{otherwise} \end{cases} \]

and \( (z)^+ = \max(z, 0) \)

We model the output of the higher priority buffers by the two-state Markov chain shown in Figure 6. Thus we term the output a "bursty binary." For simplicity, we analyze the case where \( N = 2 \) and indicate the general result at the conclusion of this appendix.

Clearly, the mean queueing delay through the high priority (H) buffer [10]

\[ W_H = \frac{\alpha_H}{2(1 - \alpha_H)} \]  (A.2)

Define

\[ p_L(i; j; S = z) = P[low priority (L) buffer contains j packets at the end of slot i and the state of H buffer is z] \]

where z is equal to zero if H buffer is empty and is equal to one otherwise.

The first-order Markov model yields the following one-step transition equations:

\[ p_L(i; j; S = 0) = p_L(i-1; j; S = 0) + \sum_{k=0}^{j} p_L(i-1; j-k+1; S = 0) \]  (A.3)

\[ + \sum_{k=0}^{i} \sum_{z=0}^{i} p_L(i-1; j-k; S = 1) \alpha_L(k)(1 - b_H) \]

\[ + \sum_{k=0}^{i} \sum_{z=0}^{i} p_L(i-1; j-k; S = 1) \alpha_H(k)(1 - b_L) \]

\[ p_L(i; j; S = 1) = p_L(i-1; j; S = 0) \alpha_H + \sum_{k=0}^{j} \sum_{z=0}^{i} p_L(i-1; j-k+1; S = 1) \alpha_L(k) b_H \]

\[ + \sum_{k=0}^{j} \sum_{z=0}^{i} p_L(i-1; j-k; S = 1) \alpha_H(k)b_H \]  (A.4)

where \( \alpha_H \) and \( b_H \) are as described in Section 4 and

\[ r(i)(j) = P[j \text{ packets arrive at L buffer during slot i}] \]  (A.5)

For a stationary solution to exist, \( \alpha_H + \alpha_L < 1 \) and then (A.3) and (A.4) hold for \( p_L(j; S = z) = \)
let 
\[ \lim_{t \to \infty} p_L^{(i)}(j; S = x). \]

Letting
\[ P_L(x; S = x) = \sum_{j=0}^{\infty} p_L(j; S = x)x^j \]  
(A.6)

and \( A_L(z) \) be the z-transform of the arrival process to L buffer,
\[ P_L(0; S = 0) = p_L(0; S = 0)A_L(z)\alpha_h \]

\[ + \frac{1}{z}[P_L(0; S = 0) - p_L(0; S = 0)]A_L(z)\alpha_h \]

\[ + P_L(x; S = 1)A_L(z)\beta_h \]  
(A.7)

\[ P_L(x; S = 1) = p_L(0; S = 0)A_L(z)\alpha_h \]

\[ + \frac{1}{z}[P_L(x; S = 0) - p_L(0; S = 0)]A_L(z)\alpha_h \]

\[ + P_L(x; S = 1)A_L(z)\beta_h \]  
(A.8)

where \( z = 1 - x \).

Solving for \( P_L(x; S = 0) \) and \( P_L(x; S = 1) \) we obtain
\[ P_L(x; S = 0) = \frac{p_L(0; S = 0)(x - 1)A_L(z)[\beta_h - A_L(z)\gamma_h]}{z - zA_L(z)\beta_h - A_L(z)\alpha_h + A_L(x)\gamma_h} \]  
(A.9)

\[ P_L(x; S = 1) = \frac{p_L(0; S = 0)(x - 1)A_L(z)\alpha_h}{z - zA_L(z)\beta_h - A_L(z)\alpha_h + A_L(x)\gamma_h} \]  
(A.10)

where \( \gamma_h = \beta_h - \alpha_h \).

From total probability we have
\[ P_L(x) = P_L(x; S = 0) + P_L(x; S = 1) \]  
(A.11)

therefore the probability generating function of the occupancy of the L buffer is
\[ P_L(x) = \frac{p_L(0; S = 0)(x - 1)A_L(z)[1 - A_L(z)\gamma_h]}{z - zA_L(z)\beta_h - A_L(z)\alpha_h + A_L(x)\gamma_h} \]  
(A.12)

It follows easily (from the fact \( P_L(1) = 1 \) and \( A_L(1) = 1 \) and the application of L'Hospital's rule) that
\[ p_L(0; S = 0) = 1 - \alpha_h - \alpha_L \]  
(A.13)

We can derive an expression for the mean number of packets in the L buffer or in service, \( N_L \), in steady-state from (A.12) and the relation \( N_L = P_L^{(1)}(1) \) and twice applying L'Hospital's rule,