Université d'Ottawa · University of Ottawa

Université d'Ottawa · University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

MASKERY, Michael

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M.Sc. (Mathematics)

GRADE - DEGREE

Mathematics

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Stochastic Stability of TCP Networks Under Random Packet Dropping Schemes

David McDonald

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

Doug Down

Victor LeBlanc

Yiqiang Zhao

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

STOCHASTIC STABILITY OF TCP NETWORKS UNDER RANDOM PACKET DROPPING SCHEMES

By
Michael Maskery, B.Sc.
June 2003

A Thesis
submitted to the School of Graduate Studies and Research
in partial fulfillment of the requirements
for the degree of
Master of Science in Mathematics[1]

[1]The M.Sc. Program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Your file Votre référence
ISBN: 0-612-90115-7
Our file Notre référence
ISBN: 0-612-90115-7

Abstract

This paper investigates the stability of TCP networks when packets are randomly dropped at bottleneck routers with a constant or near-constant probability. Analysis of a previously developed system of stochastic differential equations leads to the proposal of a new router algorithm, RWFD, which drops packets with a nearly constant probability. Stability is then investigated for a single TCP connection when this probability is constant. The connection is viewed on a new time scale and modelled as a general state-space Markov chain. Ergodic theory and Foster-Lyapunov drift conditions are employed to show that the Markov chain converges to a steady-state distribution. Stability for near-constant loss probabilities is also considered. The results are extended through the Law of Large Numbers to conclude that constant drop probabilities may cause large TCP networks to converge to a known fixed point. Simulation verifies that RWFD is similarly well behaved, while automatically adapting to network conditions.

Acknowledgements

I would like to thank my supervisor Professor David McDonald for all his guidance and advice. I would also like to thank Michel Ouellette and Alan Chapman from Nortel Networks for their many helpful discussions, Laurent Fournié for his help and enthusiasm, and Oliver Yang for his assistance in obtaining the Opnet simulator.

I would like to thank both the Ontario Graduate Scholarship for Science and Technology program and the Ottawa-Carleton Institute of Mathematics and Statistics for their funding throughout my research.

Dedication

To Caragh. Thank you for your patience and support.

Contents

List of Figures

Chapter 1

Introduction

The growth of the World Wide Web has presented mathematicians and engineers alike with many challenges. Not only is there an ever increasing number of computer users, but the amount of data being transferred per user across networks, thanks to video, voice, and real-time interactive software, is much larger than ever before. The increased demand for bandwidth is pushing the limits of the dominant protocol for data transfer, TCP (Transport Control Protocol), which was developed 20 years ago without any consideration for today's traffic levels. It is no longer feasible, as it was just a few years ago, to solve the problem of network congestion by simply building networks with excess capacity. The search is on for ways to use available network capacity in a more efficient, economical fashion.

Although TCP is able to send information reliably under almost any network conditions, it does so with a highly variable transmission rate. This is because of the dual requirement that TCP must maximize its use of bandwidth but still react quickly to network congestion. When many TCP connections share network resources, these variable transmission rates can lead to large fluctuations in queue sizes and transmission delays. Needless to say, such fluctuations are undesirable and must be regulated.

To control a system such as a TCP network, mathematical models must first be created and analyzed. Conditions under which the network is well-behaved must be defined, and strategies to ensure that these conditions are met must be developed.

This thesis works toward these goals.

The problem of TCP network stability, in which the amount of data in the network, and the mean data transfer rate are constant, is approached in the following manner. The mathematical model developed in [1] is used to model a single TCP source as a Markov chain. With the general state-space Markov chain theory developed in [2], it can then be shown that, under certain conditions, this source settles to a steady-state. This result is then extended to conjecture that a large network of TCP sources will also reach steady state, under the right conditions. A parallel theme is the development of a new TCP control protocol, called RWFD, which may be implemented in practice to approach the conditions necessary for stability.

Before reaching the main body of theory, a review of the technical details surrounding TCP behavior is in order. Complete and up-to-date information can be found online at www.ietf.org, by browsing the RFC pages.

1.1 A Brief Review of TCP

The TCP/IP set of protocols is currently the most common method for sending information over the internet. A computer uses TCP to send packets of data to a receiver over a data link of unknown capacity and delay in such a way that all the data is received in order, and the use of the available bandwidth is maximized.

To ensure that all the data is received in order, the receiver must provide feedback by acknowledging data packets received. This is done by attaching a sequence number to each packet sent by the source. The source assigns numbers in increasing order to each packet it sends into the network, and writes them into the current packet header. The receiver reads this sequence number and sends a message to the source, specifying the last packet that has been received in order and without interruption. This implies that if there is a gap in the sequence of packets received, then the receiver will send a series of identical acknowledgements until the problem is resolved. If the source receives such duplicate acknowledgements, it concludes that data has been lost and retransmits the missing information.

To utilize connection bandwidth effectively, TCP sources use a congestion window

(often referred to as just the window). This is a state variable in each TCP source that specifies how many unacknowledged packets of data are allowed in the network at any given time. As data is acknowledged, the source gains information about the network between itself and the receiver, and may make appropriate decisions to dynamically adjust the congestion window size.

Although many versions of TCP exist, we will concentrate exclusively on the case where sources implement TCP Reno, and receivers acknowledge every packet received. Assume that a TCP source has a large block of data to send over a network. In this case, the TCP protocol for a single connection functions in the following manner.

1. A connection is established between two computers. The sender is aware of both the address of the receiver and the location in memory of the data to be sent. The receiver is aware that incoming data will arrive from the sender and allocates memory for it. No other part of the network is aware of the connection.

2. The source enters the slow start phase. The congestion window size is set to a small number, and a few packets are transmitted. The window size is then increased by one each time an acknowledgement is received. Upon such a receipt, the source may transmit two packets- one packet because the arrival of an acknowledgment signals that there is room in the network, and another packet because the window size has been increased. The end result is that the data transfer rate increases exponentially with time. If a packet loss is detected, the source halves its window size and immediately enters congestion avoidance.

3. When a certain predetermined window size is reached, or when a loss is detected, the source leaves slow start and enters the congestion avoidance phase. In this mode of operation, the window size increases at a much slower rate. If the window size is n, then it will increase to $n + 1$ only after n acknowledgements are received. The result is an approximately linear increase in the window size, since it takes an approximately constant time (one round-trip time) for an entire window worth of acknowledgements to return to the source, regardless of the actual window size. Again, if a packet loss is detected, the source halves its

window size, retransmits the missing information, and continues in congestion avoidance.

4. If no acknowledgements are received for a long time (about one half to two seconds, depending on the implementation) the source resets and enters slow start.

The behavior described above represents a very simplified outline of TCP. There are other important features. The TCP source monitors the round-trip time, which is the time from the transmission of a packet to the receipt of its acknowledgment. This is done so that timeouts can be detected. Another algorithm, Fast Retransmit/Fast Recovery, retransmits packets when a loss occurs in a manner that does not cause the whole scheme to collapse. There is also a (very large) maximum window size that can be reached. However, these features are relatively invisible from a modelling standpoint, and it is more useful to simply assume that they are working properly.

In addition to this, for large data transfers, sources spend very little time in slow start. This is because the slow start phase lasts for less than $\log_2(ssthresh)$ round-trip times, where *ssthresh* is certainly less than 128 for any realistic network. In addition, slow start only occurs at the beginning of a connection or after a timeout, and in a properly designed network, timeouts must be relatively rare events. Thus the connection is likely to be governed by congestion avoidance for the majority of the data transfer and it is justifiable to only model the congestion avoidance phase when considering long-lived connections.

1.2 A Brief Review of RED

When many TCP connections share the same link, each will continue to increase its network usage according to the rules above. As the number of packets in the network increases, bottlenecks develop at points where the outgoing data link cannot transmit packets as fast as they are arriving. These points are typically routers, and these routers queue excess packets until they can be sent.

However, even the routers have a finite amount of storage space, and if the TCP

sources continue to increase their window size, the router queues will overflow and cause packets to be lost. When a queue starts to overflow, every source sending packets through the offending router will experience a packet loss. It follows that all the TCP window sizes will be halved virtually simultaneously. This causes a dramatic drop in the number of packets in the network, which temporarily reduces the queue size, but the effect will repeat as the window sizes increase again. This leads to a phenomenon known as global synchronization. Since losses occur at the same moment for all sources, the TCP window sizes will rise and fall at the same time, and eventually all the window sizes will become almost identical. Since the number of packets in the network is related to the sum of window sizes, it will exhibit large periodic fluctuations, with amplitude approaching the number of sources times the amplitude of the individual window size fluctuations.

Such behavior causes many undesirable effects, such as large queueing delay variation, low bandwidth utilization, and even fairness issues between connections. To alleviate such problems, a packet dropping policy known as RED (Random Early Detection) was proposed. (See [8] and [9].) Routers implementing RED policies randomly drop any incoming packet with a certain probability. This probability is a linearly increasing function of the queue size, and typically ranges from zero for an empty queue, to some value less than ten percent at the maximum queue size. The goal of RED is to spread out packet losses over time, so that individual TCP connections reduce their window size before the queue overflows. Properly tuned RED algorithms have been shown to increase network performance by avoiding the synchronization effects associated with regular Tail-drop routers (see [8] and [5]). The queue sizes can potentially become constant at a moderate level, and the associated variations in delay can be minimized.

The parameters of RED are given by p_{max}, p_{min}, Q_{max}, and Q_{min}. The router implements RED by estimating the queue size, $Q(t)$, at time t. If $Q_{min} < Q(t) < Q_{max}$, then incoming packets are dropped randomly with probability

$$k(t) = p_{min} + \frac{p_{max} - p_{min}}{Q_{max} - Q_{min}}(Q(t) - Q_{min}).$$

If $Q(t) < Q_{min}$ then the packet is accepted into the queue, if $Q(t) > Q_{max}$ then the

packet is dropped. Typically, p_{min} is zero. The slope of the RED function is defined to be dk/dQ, and is equal to $(p_{max} - p_{min})/(Q_{max} - Q_{min})$.

For a given network, there is often a very small range of effective RED functions that a router can employ. The problem is that large slopes may lead to the same synchronization effects as Tail-drop routers, while small slopes are too specific to control a network with changing conditions. Finding the correct parameters and adapting them to satisfy changing network conditions is still a problem without satisfactory solution. A large number of competing solutions, either improving on RED or offering alternative algorithms, have been put forward, but a clear winner has yet to emerge.

Chapter 2

A Mean-Field Model for TCP

The random packet dropping philosophy is a good one, but its implementation is difficult. One must first understand how TCP networks behave before creating a proper control strategy. Naturally, the first step towards this understanding is the creation of an appropriate model. In this section we describe a model for a large number of TCP sources sharing a single bottleneck router. This model was developed in [1], and a similar, simplified model appears in [5].

2.1 The Dynamics of a Single TCP Source

Consider just one TCP source executing congestion avoidance (label this source n out of a possible N sources). Its state at any time t can be described in terms of both the size of the TCP congestion window and the round-trip time experienced by packets sent into the network. The congestion window increases by one packet per round-trip time, except when a packet loss is detected. This is because it takes approximately one round-trip time for an entire window size worth of acknowledgements to return, triggering an increment in the window size. When a loss is detected, the congestion window drops to half its current value.

Let $W_n(t)$ and $R_n(t)$ represent the window size and round-trip time, respectively, of source n at time t. These two variables will be enough to specify the state of a TCP source in congestion avoidance.

According to (3.1) of [1], if the window size is regarded as being able to take on a continuous range of values, the evolution of the window size of source n is described by the following stochastic differential equation.

$$dW_n(t) = \frac{1}{R_n(t)} \cdot (1 - \chi_{S_n(t)})dt - \frac{W_n(t^-)}{2} dN_n(\Lambda_n(t)) \tag{2.1}$$

In this equation, $\chi_{S_n(t)} = 1$ when source n is in the fast retransmit/fast recovery phase, and zero otherwise. During this phase, the window size does not follow a linear increase, and no losses can be detected. Since the fast retransmit/fast recovery periods are relatively short, we may let $\chi_{S_n(t)} \equiv 0$, ignoring the term without significantly changing the dynamics. t^- represents a time just before time t. As in [1], $N_n(t)$ is a Poisson process with intensity 1, and

$$\Lambda_n(t) = \int_0^t \frac{W_n(s - R_n(s))}{R_n(s - R_n(s))} F(Q(s - R_n(s)))ds \tag{2.2}$$

is the compensator for the process. Here, $F(Q(t))$ is the packet drop probability at a router serving the source as a function of its queue size, as is the case with RED. A more general drop policy would replace this factor with $k(t)$, so that the drop probability need not depend explicitly on the current queue size.

$N_n(\Lambda_n(t))$ is then a nonhomogeneous Poisson process with variable intensity

$$\lambda_n(t) = \frac{W_n(t - R_n(t))}{R_n(t - R_n(t))} F(Q(t - R_n(t))) \tag{2.3}$$

(See, for example, [3], section 5.5.4.)

It is important to note that we define the round-trip time by

$$R_n(t) = T + \frac{Q(t - R_n(t))}{L}. \tag{2.4}$$

Here, T represents the round-trip transmission delay outside of the queue, and L represents the network transmission rate in packets per second, so that, by Little's formula, $Q(t)/L$ is the queueing delay of packets arriving to the router at time t. There is a subtle but important modelling distinction here. (2.4) defines the round-trip time based on the past queue size, so that there is an implicit delay built in to the model. We can interpret (2.4) in the following way. Recall that a TCP source transmits a new

packet whenever an acknowledgement arrives. If an acknowledgement arrives at time t, then its associated round-trip time, $R_n(t)$, is based on the delay experienced by the packet that generated it. This packet reached the router approximately $R_n(t)$ time units ago, so its queueing delay was $Q(t - R_n(t))/L$. $R_n(t)$ can thus be viewed as the round-trip time experienced by packets *arriving* to source n at time t, instead of the round-trip time that will be experienced by packets leaving the source at time t. We may imagine that the source writes this value into the new packet being transmitted at time t.

2.2 Interaction of Sources at a Bottleneck Router

If a number of TCP sources send packets through a common queue, such as the one occurring at a bottleneck router, the dynamics of the sources become coupled together. For example, if a certain source shares a queue with other sources having large window sizes, then there will be many packets in the router queue. This will lead to longer round-trip times, and a higher loss probability if the router is implementing a RED-like drop policy.

Assume that the round-trip times of all the sources are the same, or that they may be replaced by an equivalent "mean" round-trip time. This round-trip time varies in time with the queue size.

The interaction of N TCP sources at the bottleneck queue implementing a RED-like drop policy leads to a queue size governed by (3.2) of [1]. This formulation uses the relative, per-source queue size $q(t) = Q(t)/N$, and relative link rate $C = L/N$. The relative queue size $q(t)$ is assumed to be fluid-like, giving

$$\begin{aligned}\frac{dq(t)}{dt} &= \frac{1}{N}\sum_{n=1}^{N}\frac{W_n(t)}{R_n(t)}(1 - F(Nq(t))) - C \\ &+ \left(\frac{1}{N}\sum_{n=1}^{N}\frac{W_n(t)}{R_n(t)}(1 - F(Nq(t))) - C\right)^{-}\chi\{Nq(t) = 0\}. \qquad (2.5)\end{aligned}$$

The second term in the equation keeps the queue from becoming negative, by adding the expressed quantity if it is negative, when the queue size is zero. For

a system of N TCP sources, we now have a stochastic system of $N+1$ differential equations. That is, there are N equations governing the window sizes, as in (2.1), and one equation governing the queue size, as in (2.5). We will now deal with the question of what happens when N becomes large, as this approach will give a deterministic result in the spirit of the Law of Large Numbers.

2.3 The Mean-Field Limit

Mean-field theory involves replacing many random interacting systems with an equivalent effective system. This approach can be applied to TCP networks in the limit-when the number of sources becomes large. Instead of considering N individual window sizes, one can re-formulate the model and consider a histogram giving the frequency of connections with a certain window size. The advantage of this is that only one equation is needed to specify all the window sizes, instead of having one equation for each source.

Following [1], (2.1) and (2.5) lead to a deterministic system of partial differential equations as the number N of TCP sources becomes large. The limiting mean-field equations take the form of (3.12) and (3.13) of [1], which are reproduced below as (2.6) and (2.7).

Let N be the number of sources executing congestion avoidance. Let $q(t)$ be the relative queue size per active connection treated as a fluid and let $p(t,w)$ be the histogram of the window sizes of connections in congestion avoidance at time t. Assume for the moment that all sources have the same round-trip time. According to [1], in the mean-field limit as the number of sources becomes large, we obtain

$$\begin{aligned}\frac{\partial p(t,w)}{\partial t} &= -\frac{1}{r(t)}\frac{\partial p(t,w)}{\partial w} \\ &\quad +\Big(4wp(t,2w) - wp(t,w)\Big)\frac{k(t-r(t))}{r(t-r(t))},\end{aligned} \tag{2.6}$$

and

$$\frac{dq(t)}{dt} = \int_0^\infty \frac{w}{r(t)} p(t,w)dw(1-k(t)) - C \qquad (2.7)$$
$$- \left(\int_0^\infty \frac{w}{r(t)} p(t,w)dw(1-k(t)) - C \right)^- \chi\{q(t)=0\}.$$

C is the link rate per source in packets per second per active source, i.e. $L = NC$, $k(t)$ is the probability that each packet is dropped and $r(t) = T + q(t - r(t))/C$ is the effective or aggregate round-trip time of all the sources with T being the common transmission time and $q(t - r(t))/C$ being the common queueing time of packets arriving to the queue one round-trip time ago.

The heuristics of (2.7) are clear. The rate of change in the relative queue size is the difference between the rate per source at which fluid arrives minus the link rate per source C. The mean arrival rate per source is the mean window size $\int_w wp(t,w)dw$ divided by the round-trip time $r(t)$ times the proportion $(1 - k(t))$ which is not discarded. The second term only keeps the queue from becoming negative.

The heuristics of (2.6) are not so complicated either. Integrating out both sides of (2.6) quickly establishes that $\int_w p(t,w)dw$ is constant; i.e. that $p(t,w)$ is a density for all times t. The left-hand side of (2.6) times dw is equal to the rate at which mass flows into and out of an infinitesimal slice of windows $(w, w + dw]$. The first term on the right-hand side times dw is due to the additive increase of the component of TCP; every round-trip time the window increases by one. In an infinitesimal period of time dt the quantity of fluid that pours into the slice is given by $p(t,w)dt/r(t)$ to first order while the quantity that pours out is $p(t, w + dw)dt/r(t)$ to first order. The difference is approximately $-\frac{dt}{r(t)}\frac{\partial p(t,w)}{\partial w}dw$. It follows that the rate at which the mass in the slice $(w, w + dw]$ is growing because of the additive increase is $-\frac{1}{r(t)}\frac{\partial p(t,w)}{\partial w}dw$. The second term is due to the multiplicative decrease. A loss one round-trip time in the past, i.e. at $t - r(t)$, among sources in the slice $2(w, w + dw]$ causes mass from the slice $2(w, w + dw]$ to be dumped into the slice $(w, w + dw]$. In an infinitesimal time interval dt the probability of such a loss is equal to dt times the transmission rate $(2w/r(t - r(t)))$ of windows in the slice $2(w, w + dw]$ times the loss probability $k(t - r(t))$. The mass dumped from the interval $2(w, w + dw]$ at time t when the

loss is detected is $p(t, 2w)2dw$. Hence the rate at which mass is added to the interval $(w, w + dw]$ is $p(t, 2w)\frac{2w}{r(t-r(t))}2dw$. Similarly a loss one round-trip time in the past among sources in the slice $(w, w + dw]$ causes mass to be subtracted from this slice. The rate at which mass is subtracted from the interval $(w, w+dw]$ is $p(t, w)\frac{w}{r(t-r(t))}dw$. The equation (2.6) describes the net effect of these flows.

2.4 A Fixed Point for the Mean-Field Model

When the drop probability $k(t)$ is well chosen, perhaps as a function of $Q(t)$ as in RED, then (2.6) and (2.7) may become constant in time; the drop probability $k(t)$ tends to a constant k, the window distribution stabilizes to a fixed distribution f_k, $q(t)$ tends to a constant q and the round-trip time, $r(t)$, tends to a constant r. For stable systems (2.6) and (2.7) become:

$$\frac{df_k(w)}{dw} = k\left(2(2w)f_k(2w) - wf_k(w)\right) \tag{2.8}$$

$$C = (1-k)\frac{1}{r}\int_w wf_k(w)dw. \tag{2.9}$$

(2.9) is simply Little's formula since the right-hand side represents the throughput as the average window size divided by the round-trip time times the proportion of packets that are not killed.

The following theorem is given in [1].

Theorem 2.1 *Let* $\Psi = \sum_{i=0}^{\infty} \frac{2^i}{\prod_{j=1}^{i}(1-4^j)}$ *(*$\Psi \approx 0.4194$*). The unique density* $f_k(w)$ *solving (2.8) is given by*

$$f_k(w) = \sum_{i=0}^{\infty} a_i \exp(-k4^i\frac{w^2}{2}) \tag{2.10}$$

$$a_0 = \sqrt{\frac{2}{\pi}}\frac{1}{\Psi}\sqrt{k}\ ;\ a_i = a_{i-1}\frac{4}{1-4^i} = a_0\frac{4^i}{\prod_{j=1}^{i}(1-4^j)}. \tag{2.11}$$

The mean window size is

$$\int_w wf_k(w)dw = \alpha\sqrt{\frac{1}{k}} \tag{2.12}$$

where $\alpha \approx 1.310$. *Combining (2.12) with (2.9) yields*

$$\frac{(1-k)^2}{k} = \left(\frac{rC}{\alpha}\right)^2. \tag{2.13}$$

This theorem gives valuable information about the fixed points, which will be used in the controller of the next chapter.

The above mean-field arguments apply even if the transmission delays vary between sources. The loss probability may still stabilize at k and the queue may still stabilize at q. Assume that the steady state above is achieved. In this case an individual source will evolve like a dynamical system uncoupled from the queue and hence other sources. If the window of source n is denoted by $W_n(t)$ at time t and the round-trip time of source n is $r_n = T_n + q/C$ then

$$W_n(t) = \int_0^t \left[\frac{ds}{r_n} - \frac{W_n(s)}{2} dN_n(\Lambda_n(s))\right],$$

where $N_n(t)$ is a Poisson process with intensity 1 and

$$\Lambda_n(t) = \int_0^t \frac{W_n(s-r_n)}{r_n} k ds$$

is the stochastic intensity for the Poisson point process of losses of connection n. (See (2.1) and (2.2).) Hence the window reductions at source n occur according to the time-changed Poisson process $N_n(\Lambda_n(t))$.

With time change $s = r_n u$:

$$\begin{aligned} W_n(t) &= \int_0^{t/r_n} \left[du - \frac{\Theta_n(u)}{2} dN_n(\Lambda_n(r_n u))\right], \text{ or,} \\ \Theta_n(s) &= \int_0^s \left[du - \frac{\Theta_n(u)}{2} d\sigma_n(u))\right] \end{aligned}$$

where $\Theta_n(s) = W_n(r_n s)$ and $\sigma_n(u)$ is a Poisson process with intensity $k\Theta_n(u-1)$. The evolution of Θ_n is therefore independent of the round-trip time and is the same for all windows. In effect all the windows evolve in the same way but at different speeds.

We conclude that the joint distribution of the window sizes and the RTTs will stabilize to a product distribution $f_k(w)g(r)$. For stable systems (2.6) and (2.7)

become

$$\begin{aligned} \frac{df_k(w)}{dw} &= k\left(2(2w)f_k(2w) - wf_k(w)\right) \\ C &= (1-k)\int_w wf_k(w)dw \int_r \frac{1}{r}g(r)dr. \end{aligned}$$

Hence everything remains unchanged if we replace r by $r_{equ} = (\int_r \frac{1}{r}g(r)dr)^{-1}$. Note that this equivalent round-trip time is calculated much as one calculates the effective resistance of resistances in parallel. This fact has been remarked in [5].

Chapter 3

RED Without Feedback Delay

There are a vast number of competing protocols designed to control the behavior of TCP networks. Many of these protocols are variants of RED, and use the control mechanism of random packet drops in an attempt to stabilize the queue size at the router. These protocols are collected under the Active Queue Management umbrella, and are described by a rainbow of acronyms such as RED, FPQ, DRED, BLUE, FRED, etc. The term active *queue* management betrays the perhaps short-sighted nature of many of these protocols, because true stability can only be achieved by managing the entire network, which consists of much more than a single queue. Such control strategies may break down when a significant portion of the data lies outside of the managed queue, as is the case when the number of connections is small compared to the number of packets that can be held in the network outside of the queue. Such a case can occur when outside transmission delays are very long.

This section presents yet another management protocol, RWFD, which randomly drops TCP packets with a probability designed to account for *all* network conditions, not just the router queue size. The algorithm, based on the mean-field model of the previous chapter, attempts to compensate for the delays in the control mechanism which lead to oscillatory behavior of the queue size. A further advantage of this protocol, as will be seen in later chapters, is that the probability of dropping incoming packets is relatively constant.

3.1 The RWFD Strategy

The factor $k(t - r(t))/r(t - r(t))$ in (2.6) captures two causes of oscillations leading to network instability. The control signal $k(t - r(t))$ is delayed by one round-trip time $r(t)$ and the system depends explicitly on the past through the denominator $r(t - r(t))$. To cancel out this delay and eliminate the associated instabilities, choose the drop probability

$$k(t) = \frac{r(t)}{r_{equ}} k_{equ}, \tag{3.1}$$

where k_{equ} satisfies the fixed-point equation (2.13),

$$\frac{(1 - k_{equ})^2}{k_{equ}} = \left(\frac{r_{equ}C}{1.31}\right)^2 \tag{3.2}$$

and $C = L/N$. r_{equ} is the effective round-trip time of all the sources if the relative queue size were to stabilize at q_{target}, the target relative queue size. r_{equ} and N must be estimated in practice. Note that if k_{equ} is small, (3.2) has approximately the same solution as

$$\frac{1}{k_{equ}} = \left(\frac{r_{equ}C}{1.31}\right)^2. \tag{3.3}$$

If the packet discard probability is chosen as above then (2.6) and (2.7) become

$$\frac{\partial p(t,w)}{\partial t} = -\frac{1}{r(t)} \frac{\partial p(t,w)}{\partial w} \tag{3.4}$$

$$+ \left(p(t,2w)\frac{2w}{r_{equ}} 2 - p(t,w)\frac{w}{r_{equ}}\right) k_{equ}, \tag{3.5}$$

and

$$\frac{dq(t)}{dt} = \int_w \frac{w}{r(t)} p(t,w) dw (1 - \frac{r(t)}{r_{equ}} k_{equ}) - C, \tag{3.6}$$

where $q(t)$ has a sticky boundary at 0, just like in (2.5). Since delays often cause system oscillations, it is hoped that eliminating the explicit delays will lead to a more stable system. If this is the case, the system above will be more likely to converge to the unique steady state of the last chapter. In steady state the relative queue size is q_{target}, the loss probability is k_{equ} and the effective round-trip time is r_{equ}.

Note that $r_{equ} = T + q_{target}/C$, so the control signal $k(t)$ satisfies

$$\begin{aligned} k(t) &= \frac{r(t)}{r_{equ}} k_{equ} \\ &= \left(1 + \frac{q(t - r(t)) - q_{target}}{C r_{equ}}\right) k_{equ} \qquad (3.7) \\ &= \left(1 + \frac{Q(t - r(t)) - Q_{target}}{L r_{equ}}\right) \left(\frac{1.31N}{L r_{equ}}\right)^2. \qquad (3.8) \end{aligned}$$

This is similar to a RED algorithm (without the moving average estimate of the queue size) with the control based on the past queue size instead of the present. The slope of the algorithm, i.e. the change in loss probability with respect to queue size, can be seen from (3.8) to be,

$$\frac{dk}{dQ} = \frac{1.716N^2}{L^3 r_{equ}^3}. \qquad (3.9)$$

In a properly-designed network, the average window size w_{avg} should be greater than five. Since the average number of packets in the network is equal to both Nw and the delay-bandwidth product Lr_{equ}, it is true that N is typically less than one fifth the size of Lr_{equ}. Consequently, the slope is typically small, and the loss probability is reasonably constant. The fact that N must be large in the mean-field model does not pose a problem here because L is scaled up with N, so the ratio of sources to delay-bandwidth may still be small in the limit.

Moreover, it has been observed through simulation that networks become less stable as the number of sources decreases, or when the link rate or delay increases. When this happens, dk/dQ decreases, leading to a more constant loss probability. As we shall see, a constant loss probability leads to system stability, so the algorithm adapts to maintain stability in the face of increasingly hostile network parameters.

This philosophy is corroborated in [5]. In that model, RED is considered stable if,

$$\frac{dk}{dQ} \leq \frac{(2N^-)^2}{L^3 (r^+)^3} (1 + \omega_g), \qquad (3.10)$$

where ω_g is considered to be small. Note that the RWFD slope satisfies this inequality if N and r_{equ} don't vary much. The linear analysis of [5] thus provides evidence of RWFD stability.

3.2 Parameter Estimation for RWFD

The above algorithm requires a knowledge of the number of connections N using the router, their equivalent round-trip time r_{equ}, and their available bandwidth L. This is not surprising since these parameters describe the most important aspects of a TCP network, and any good controller must have quantitative information on the system it wishes to control. It should be emphasized that knowledge of these key network parameters is essential to any active queue management scheme. The topology of the network is much less important (in simple models) because TCP connections are ignorant of the actual path their data takes through the network.

The round-trip time estimate, r_{equ}, is updated every time a packet is dropped. Since lost packets are retransmitted by a source approximately one round-trip time after they are dropped, we can estimate r_{equ} by storing times at which packets are dropped and measuring the time taken for the retransmitted packets to appear at the router. If there is no timeout then this gives a round-trip time sample for the studied source. We call R^j the measured round-trip time of the j^{th} dropped packet. To be more accurate, R^j is the time between the third packet following the j^{th} dropped packet and the retransmitted packet (since the source only retransmits a packet after three duplicate acknowledgments). But it is easy to compensate for this by delaying the start of our round-trip time timer until the receipt of the third packet (from the same source) after the dropped packet.

To implement the estimate, we identify each connection by reading the source address and/or the destination address of each dropped packet and creating a hash value. This value is stored in a table along with the sequence number of the dropped packet and the time of the drop. We must then calculate the hash value of the source address and/or the destination address of every incoming packet. If the value matches a table entry, the sequence number is compared with the one stored in the table in order to identify the retransmitted packet. Once the retransmitted packet arrives, the difference in the current time and the stored drop time becomes a sampled round-trip time value. An average is taken to keep a running estimate of r_{equ}, but this must be done carefully, as explained below.

The probability of picking a packet from source n is proportional to the transmission rate, where the transmission rate is proportional to the window size and inversely proportional to the round-trip time for packets from this source. Since the window distributions of sources in equilibrium are equal with mean w_{equ}, it follows that the probability of picking a packet from source n whose round-trip time is r_n is roughly

$$P(R^j = r_n) = \frac{w_{equ}/r_n}{\sum_{i=1}^{N} w_{equ}/r_i} = \frac{r_{eff}}{N r_n} \tag{3.11}$$

where

$$r_{eff} = \left(\frac{1}{N} \sum_{i=1}^{N} \frac{1}{r^i} \right)^{-1} \rightarrow r_{equ} \text{ as } N \rightarrow \infty.$$

The expected value of R^j is $\sum_{i=1}^{N} r_i (r_{eff}/N r_i) = r_{eff}$.

Fluctuations in the queue size due to randomness can affect the evaluation of the effective round-trip time. This is because r_{eff} is supposed to represent the round-trip time when $q(t) = q_{target}$, but the sample measurements are taken at arbitrary values of $q(t)$. We can compensate for this since we know the queue size when the sample is taken, and hence its difference from the target queue size. We re-center our measurements R^j as $R^j_{eff} = R^j + (q_{target} - q^j)/C$ where q^j is the relative queue size when we measured R^j. In other words we measure the round-trip time as it would have been if the queue size were on target. The R^j_{eff} now give an independent and identically distributed sample of the round-trip times of the sources picked for discard if the queue was stabilized at the target. The exponentially weighted moving average $\hat{r}^j_{equ} = \alpha \hat{r}^{j-1}_{equ} + (1 - \alpha) R^j_{eff}$ will converge to a distribution centered at r_{eff} so we can use $\hat{r}^j_{equ}$ as our estimator for r_{equ} based on the first j measurements.

The other network parameter required is the number N of sources executing congestion avoidance, and this can be estimated using the FPQ estimate given in [7]. The technique requires a regular sampling of arriving packets. A hash value of the destination address in the packet header of the sampled packets must be calculated and the corresponding entry in a hash table must be set. The estimate for N is based on the number of entries set in the hash table. An exponentially weighted moving average of the estimate is kept and used to determine the drop probability as in (3.8).

There is one other network parameter determining the behavior of the algorithm: the bandwidth L available to TCP sources in congestion avoidance. In principle the bandwidth is a known physical characteristic, but in fact it is reduced by other traffic types that do not respond to loss by reducing their transmission rates. UDP traffic is one example. Short bursts of TCP traffic called web mice also don't respond to losses because the burst is over before the source can detect a loss. It might be best to avoid dropping packets from such sources. This could only be done if the header of each packet is inspected. If this is possible it would be beneficial to correct the link rate L by subtracting off that proportion of the bandwidth taken away by these misbehaving sources (Once these sources are identified, they may also be penalized).

The above algorithm is quite robust and will behave properly even if the network parameters r_{equ}, N and L are poorly estimated. The only effect will be that the queue will stabilize at some value different from the target queue. The exact value is not of paramount importance. In fact, we may implement a high-level controller to adjust the target queue length in order to keep the rate of congestion notifications or the proportion of connections in timeout within a preferred envelope. For instance Algorithm 1 in [6] provides a load adapting mechanism which keeps the packet loss probability close to a target by adjusting the target queue length. We adjust q_{target} periodically according to the control loop

$$q_{target} = q_{target} + \beta(k(t) - k_0)$$

where β is the constant gain of our control and k_0 is some specified target drop probability (2% perhaps). We can also design the algorithm to avoid an empty queue and hence keep the utilization high. We can accomplish this with a no-drop region. We simply cease dropping cells if $Q(t)$ falls below some minimum threshold.

3.3 Commentary

The RWFD algorithm is one among many competing algorithms. Although modifications such as the load adapting mechanism and no drop region complicate matters, the algorithm exhibits two important features. First, and most importantly, the loss

probability varies only slightly as the queue size changes, since the slope in (3.9) is small. Second, the algorithm specifies both an equilibrium queue size and an equilibrium loss probability, allowing one to ensure that the algorithm effectively controls the TCP sources while maintaining a positive queue size.

Neither of these features is remarkable on its own, but, taken together, they ensure that an almost constant packet loss probability is imposed on the sources. Traditional RED algorithms are also able to achieve this result, but they must be tuned very precisely for the current network conditions. RED Without Feedback Delay essentially provides an automatic method for this tuning by estimating the network conditions and applying the appropriate control.

In the following chapters, we shall see how imposing a constant loss probability on a large number of TCP sources appears to lead to a steady-state distribution of window sizes. Such a steady state implies a constant rate of packet arrivals to the bottleneck queue, and this leads to a stable queue. (The near-constant random dropping mechanism provided by RWFD should similarly cause the network to approach a steady state, and comments will periodically be made to explore this assertion.)

Chapter 4

Further Analysis of a Single Source

A TCP source acting in congestion avoidance adjusts its window size according to the additive increase, multiplicative decrease behavior described in Chapters 1 and 2. This process can be described in real time by (2.1) and (2.2), and by considering the interaction of a large number of sources, one is led to the mean-field model (2.6), and (2.7).

This method provides a valid model for TCP sources executing congestion avoidance and interacting through a bottleneck queue. However, it is not immediately clear whether or not the resulting set of differential equations exhibit stability, in the sense that they converge to a steady state. One might attempt to use the methods of Lyapunov to check for such stability, but these methods are very difficult to apply to nonlocal partial differential equations incorporating delay.

A more illuminating approach is to first consider the stochastic equations governing a single congestion window. Stochastic stability can be studied for the single window, and conclusions about stability of the entire system can then be drawn. In fact it is possible, for a single TCP source, to re-write these equations on a time scale that counts round-trip times. This is the primary goal of this chapter, as it will simplify matters significantly by ridding the equations of variable round-trip times. Then, since a steady state is time-invariant, the steady state window size distribution (if it exists) will be the same regardless of the time scale. That is, the steady-state derived from the time-changed process will be the same as the steady state for the

original process.

4.1 A Canonical Equation for the Window Sizes

For TCP connection n, define $\tau_n^{-1}(t)$ to be an invertible mapping which transforms the time scale from one measured in seconds to one measured in the number of round-trip times that have elapsed for the connection since time $t = 0$. In other words, $\tau_n(s)$ maps "rtt" time to real time, and hence satisfies the difference equation,

$$\tau_n(s-1) = \tau_n(s) - R_n(\tau_n(s)),\ \tau_n(0) = 0. \tag{4.1}$$

(4.1) simply expresses the fact that the counting of any round-trip time is completed exactly one round-trip time before the next is. For a given function $R_n(t)$, $\tau_n^{-1}(t)$ can be at least approximately constructed by first taking $t_0 = 0$, and then recursively defining t_i by solving $R_n(t_i) = t_i - t_{i-1}$. Setting $\tau_n^{-1}(t_i) = i$ for each i and defining $\tau_n^{-1}(t)$ to be the series of line segments connecting these points creates an approximate function $\tau_n^{-1}(t)$ that exactly satisfies (4.1) at each t_i. Since the t_i are only one round-trip time apart, this construction is then a good approximation to the definition of $\tau_n^{-1}(t)$.

On this new time scale, the window size increases at a rate of one per unit time, and a delay of one round-trip time in real time becomes a delay of one time unit in "rtt" time. For a TCP window $X_n(t)$, a canonical system of equations similar to (2.1) and (2.2) can thus be written, in transformed time, as

$$dX_n(t) = dt - \frac{X_n(t^-)}{2} dN_n(\Lambda_n(t)) \tag{4.2}$$

$$\Lambda_n(t) = \int_0^t X_n(s-1)k(\tau_n(s-1))ds. \tag{4.3}$$

Notice that (4.3) may be written, through the change of variables $u = s - 1$, as

$$\Lambda_n(t) = \int_{-1}^{t-1} X_n(u)k(\tau_n(u))du. \tag{4.4}$$

The time-changed dynamics bear several similarities to the original dynamics, (2.1) and (2.2), while suppressing the role of $R_n(t)$. $X_n(t)$ represents the window

size as a function of rtt time, $N_n(t)$ is a rate-one Poisson process, and $\Lambda_n(t)$ is the stochastic compensator, with an integrand delayed one round-trip time in the past, all as before. However, the term $R_n(t)$ is normalized to one on this time scale, so it no longer appears in the denominator. Also notice that the router drop policy $k(t)$ is unchanged, since it is evaluated at the same real time as before.

The system of equations (4.2) and (4.3) is almost simple enough to analyze using stochastic methods. The central mathematical topic of this thesis is demonstrating the stochastic stability of this system under the assumption of a constant, or near-constant, drop probability $k(t)$.

4.2 Equivalence of Systems under a Time Change

In this section, we verify that the canonical equations for window size evolution, (4.2) and (4.3), are approximately a time-changed version of the real-time equations, presented in [1] and given here as (2.1) and (2.2). Consequently, stochastic stability of the canonical system will coincide with stochastic stability of the real-time system, at least for a single TCP connection.

To proceed, assume that the time change τ_n is invertible. If $W_n(t)$ represents the window size in real time, and $X_n(t)$ represents the window size in rtt-time, then

$$W_n(t) = X_n(\tau_n^{-1}(t)),$$

or,

$$X_n(s) = W_n(\tau_n(s)).$$

Now use (4.2) to write

$$\begin{aligned} dW_n(t) &= dX_n(\tau_n^{-1}(t)) \\ &= \dot{\tau}_n^{-1}(t)dt - \frac{X_n(\tau_n^{-1}(t^-))}{2} dN_n(\Lambda_n(\tau_n^{-1}(t))) \\ &= \dot{\tau}_n^{-1}(t)dt - \frac{W_n(t^-)}{2} dN_n(\Lambda_n(\tau_n^{-1}(t))). \end{aligned} \tag{4.5}$$

Using (4.4) and applying the change of variables $v = \tau_n(u)$, we may write (4.3) as,

$$\begin{aligned}\Lambda_n(\tau_n^{-1}(t)) &= \int_{-1}^{\tau_n^{-1}(t)-1} X_n(u)k(\tau_n(u))du \\ &= \int_{\tau_n(-1)}^{\tau_n(\tau_n^{-1}(t)-1)} X_n(\tau_n^{-1}(v))k(\tau_n(\tau_n^{-1}(v)))\dot{\tau}_n^{-1}(v)dv \\ &= \int_{-R_n(0)}^{t-R_n(t)} W_n(v)k(v)\dot{\tau}_n^{-1}(v)dv. \end{aligned} \tag{4.6}$$

The upper and lower limits of integration above were calculated, using (4.1), as

$$\begin{aligned}\tau_n(\tau_n^{-1}(t)-1) &= \tau_n(\tau_n^{-1}(t)) - R_n(\tau_n(\tau_n^{-1}(t))) \\ &= t - R_n(t),\end{aligned}$$

and

$$\begin{aligned}\tau_n(-1) &= \tau_n(0) - R_n(\tau_n(0)) \\ &= -R_n(0).\end{aligned}$$

To find $\dot{\tau}_n^{-1}(s)$, recall from elementary calculus that,

$$\dot{\tau}_n^{-1}(s) = \frac{1}{\dot{\tau}_n(\tau_n^{-1}(s))}.$$

We may estimate $\dot{\tau}_n(t)$ by

$$\begin{aligned}\dot{\tau}_n(t) &\approx \frac{\tau_n(t) - \tau_n(t-1)}{1} \\ &= R_n(\tau_n(t))\end{aligned}$$

This is a fair estimate because $\tau_n(t) - \tau_n(t-1)$ is a small interval (one round-trip time) so $R_n(\cdot)$ is approximately constant over the interval.

We therefore obtain the desired derivative,

$$\begin{aligned}\dot{\tau}_n^{-1}(s) &= \frac{1}{R_n(\tau_n(\tau_n^{-1}(s)))} \\ &= \frac{1}{R_n(s)}.\end{aligned} \tag{4.7}$$

Note that since $R_n(t) > 0 \; \forall \; t$, $\tau_n(t)$ is indeed invertible.

Using the approximation given by (4.7), (4.5) and (4.6) may be written as,

$$dW_n(t) = \frac{dt}{R_n(t)} - \frac{W_n(t^-)}{2} dN_n(\Lambda_n(\tau_n^{-1}(t))), \tag{4.8}$$

and,

$$\Lambda_n(\tau_n^{-1}(t)) = \int_{-R_n(0)}^{t-R_n(t)} \frac{W_n(v)}{R_n(v)} k(v) dv. \tag{4.9}$$

This pair of equations is similar to (2.1) and (2.2), except that delay is taken into account in the limits of integration instead of in the integrand. In fact, (4.8) and (4.9) could be used as an alternative model to the one found in [1] with few changes. In this way, it can be said that the canonical system approximates the real-time system.

It is desirable to have complete agreement with previous models. For completeness then, the rest of this section shows how the canonical system approximates the exact model presented in [1].

Recall from (2.4) that, for a single bottleneck queue, $R_n(t)$ satisfies

$$R_n(t) = T + \frac{Q(t - R_n(t))}{L}.$$

Differentiating both sides with respect to t yields

$$\dot{R}_n(t) = \frac{\dot{Q}(t - R_n(t))}{L}(1 - \dot{R}_n(t)),$$

which implies both

$$\dot{R}_n(t) = \frac{\dot{Q}(t - R_n(t))}{L + \dot{Q}(t - R_n(t))}, \tag{4.10}$$

and

$$1 - \dot{R}_n(t) = \frac{L}{L + \dot{Q}(t - R_n(t))}. \tag{4.11}$$

Now, $\dot{Q}(s) \geq -L$ for all values of s. This can be seen either as a direct mathematical consequence of (2.6), and (2.7), or by simply recalling that L is the packet service rate at the router, so that the rate of queue length decrease is limited by L. Using this fact in (4.10), one arrives at the conclusion that $\dot{R}_n(t) < 1$.

This has several interesting consequences which will be used below. First, it ensures that (4.11) is always positive. Second, it fulfills the hypothesis of the following lemma.

Lemma 4.1 *If $R_n(t)$ is positive and continuous with $\dot{R}_n(t) < 1$ for all values of t, then the equation*

$$s - R_n(s) = t - R_n(t)$$

has the unique solution $s = t$ for any t.

Proof Let $f(t) = t - R_n(t)$. Then

$$\dot{f}(t) = 1 - \dot{R}_n(t) > 0.$$

Thus f(t) is everywhere increasing, so clearly $f(s) = f(t)$ if and only if $s = t$. ∎

This result is an important one because it allows another change of variables to be made in (4.9). Let $s - R_n(s) = v$. Then the limits of integration in (4.9) are the solutions of,

$$s - R_n(s) = 0 - R_n(0),$$

and,

$$s - R_n(s) = t - R_n(t).$$

That is, the limits of integration are 0 and t.

This yields, by Lemma 4.1,

$$\Lambda_n(\tau_n^{-1}(t)) = \int_0^t \frac{W_n(s - R_n(s))}{R_n(s - R_n(s))} k(s - R_n(s))(1 - \dot{R}_n(s))ds. \tag{4.12}$$

Recent developments, (see [10]), have shown that the model in [1] erroneously omitted the factor $1 - \dot{R}_n(s)$, and (2.2) should actually take the form

$$\Lambda_n(t) = \int_0^t \frac{W_n(s - R_n(s))}{R_n(s - R_n(s))} F(Q(s - R_n(s)))(1 - \dot{R}_n(s))ds. \tag{4.13}$$

In light of this correction, it is clear that the system described by (4.8) and (4.12) is equivalent to the system described by (2.1) and (2.2). The term $\dot{R}_n(s)$ is small, so the correction is minor, and disappears in steady state by definition.

Under a time change, therefore, the canonical system is equivalent to the original system given in [1]. Thus the stability of the two systems should coincide, and we may gain all the necessary information about stability by studying the canonical system.

4.3 Constant and Varying Loss Probabilities

The compensator equation, (4.3), immediately simplifies in the case where the loss probability is constant. Since $k(\cdot) \equiv k$, the equation becomes

$$\Lambda_n(t) = \int_0^t X_n(s-1)k ds. \tag{4.14}$$

The constant loss-probability case will be important for the remainder of the thesis, since it allows a Markov chain to be constructed for a source's window size evolution. Stability theory is well-developed for Markov chains, so the goal is to demonstrate stochastic stability for the system described by (4.2) and (4.14). If this can be done, then the equivalence of systems above will imply stability for the real-time system described by (2.1) and (2.2), for the special case where the packet loss probability is constant. Establishing such stability is the focus of Chapter 6.

If a constant loss probability k leads to stability, then it is possible, although not guaranteed, that a near-constant loss probability $k(t)$ may also lead to stability. An intuitive rationale for this exploits the fact that packets are dropped randomly by the router. The randomness in packet dropping results in a natural fluctuation of the loss rate even when the probability is constant. If these random fluctuations do not affect stability, then minor deterministic fluctuations may also be acceptable. On the other hand, $k(t)$ may well vary in such a way that resonance and positive feedback are introduced into the system, leading to instability even when the variation is small.

A complete analysis for variable loss probabilities is beyond the scope of this thesis, since $k(t)$ will in general depend on multiple TCP window sizes. This makes it impossible to view one source in isolation, rendering the single-source Markov chain approach presented in the next chapter ineffective. It should be noted that although the Markov chain formulation requires a constant loss probability, the stability argument does not. In chapter 6 we will essentially show that two copies of the window evolution process can be coupled on a small set. The initial conditions die away and the two copies become the same from some time on. This coupling phenomenon leads to a steady-state window distribution when the loss probability is constant, but it is general enough to suggest that processes still couple when the loss probability varies.

In this case the window distribution would not reach a steady state, but copies of the process should still couple and become dominated by the loss probability as time goes on. This leads a time-varying form of stochastic stability, in the sense that variation in the window size will be driven only by variation in the loss probability.

4.4 Commentary

The equivalence of the real-time window size equations and the time-transformed window size equations rests on a number of assumptions and approximations. This is troubling, since these assumptions may not hold in the most general of cases. However, there is no reason to prefer a time scale of seconds when viewing the behavior of TCP connections. Indeed, the more natural time scale is the rtt scale.

One can therefore adopt the viewpoint that the actual dynamics of a TCP window size under a constant loss probability are given by (4.2) and (4.14), and that the real-time equations, (2.1) and (2.2), are the approximations. Indeed, it is possible to construct a model from first principles that yields the time-changed equations directly. In either case, the time-changed equations allow us to construct a Markov chain, and this mathematical convenience alone is sufficient reason for their use in the remaining chapters.

Chapter 5

A General State-Space Markov Chain

In this section, the TCP window size equations are re-formulated as a general state-space Markov chain so that stochastic stability arguments can be brought to bear. General state-space theory and the subsequent stability arguments are developed in [2], and the reader is referred there for a complete treatment.

Before defining a Markov chain for the process described by (4.2) and (4.14), we must further specify the behavior of the window size in time. Let $X(t)$ represent the window size of a generic connection at time t.

Lemma 5.1 *The solution to (4.2) is given by:*

$$X(t) = X(i)2^{-(N(\Lambda(t))-N(\Lambda(i)))} + \int_i^t 2^{-(N(\Lambda(t))-N(\Lambda(s)))}ds \tag{5.1}$$

The lemma can be proven by direct calculation. A proof by induction is given here.

Proof (4.2) means precisely,

$$\int_i^t dX(s) = \int_i^t ds - \int_i^t \frac{X(s)}{2} dN(\Lambda(s)) \tag{5.2}$$

$$\text{or, } X(t) = X(i) + (t-i) - \int_i^t \frac{X(s)}{2} dN(\Lambda(s)) \tag{5.3}$$

For any $t \geq i$, let $N(\Lambda(t)) - N(\Lambda(i)) = 1$. This implies there is a jump at some time u so that,

$$dN(\Lambda(u)) = 1,$$

and,

$$N(\Lambda(t)) - N(\Lambda(s)) = \begin{cases} 1, & i \leq s < u \\ 0, & u \leq s < t. \end{cases}$$

(5.1) then gives,

$$\begin{aligned} X(t) &= \frac{X(i)}{2} + \int_i^u 2^{-1} ds + \int_u^t 2^{-0} ds \\ &= \frac{X(i)}{2} + \frac{u-i}{2} + t - u \\ &= X(i) + (t-i) - \frac{X(i) + (u-i)}{2}. \end{aligned}$$

(5.3) gives,

$$X(t) = X(i) + (t-i) - \frac{X(u)}{2}.$$

Since there are no jumps in the interval (i, u), $X(t)$ increases linearly giving,

$$X(u) = X(i) + (u-i).$$

Therefore (5.1) and (5.3) are equivalent for $N(\Lambda(t)) - N(\Lambda(i)) = 1$, so the lemma holds in this case.

Now assume the lemma also holds for $N(\Lambda(t)) - N(\Lambda(i)) = k$. That is, for any such $t \geq i$,

$$X(t) = X(i) + (t-i) - \int_i^t \frac{X(s)}{2} dN(\Lambda(s)) = X(i)2^{-k} + \int_i^t 2^{-(N(\Lambda(t)) - N(\Lambda(s)))} ds. \quad (5.4)$$

Assume $N(\Lambda(t)) - N(\Lambda(i)) = k + 1$. Then there exists a time $r < t$ such that $N(\Lambda(r)) - N(\Lambda(i)) = k$. By (5.4),

$$X(r) = X(i)2^{-k} + \int_i^r 2^{-(N(\Lambda(r)) - N(\Lambda(s)))} ds. \quad (5.5)$$

Now observe that $N(\Lambda(t)) - N(\Lambda(r)) = 1$, so we can use the single-jump result, along with (5.5) to write,

$$\begin{aligned}
X(t) &= X(r) + (t-r) - \int_r^t \frac{X(s)}{2} dN(\Lambda(s)) \\
&= X(r)2^{-1} + \int_r^t 2^{-(N(\Lambda(t))-N(\Lambda(s)))} ds \\
&= X(i)2^{-(k+1)} + 2^{-1}\int_i^r 2^{-(N(\Lambda(r))-N(\Lambda(s)))} ds + \int_r^t 2^{-(N(\Lambda(t))-N(\Lambda(s)))} ds \\
&= X(i)2^{-(k+1)} + 2^{-(N(\Lambda(t))-N(\Lambda(r)))}\int_i^r 2^{-(N(\Lambda(r))-N(\Lambda(s)))} ds \\
&\quad + \int_r^t 2^{-(N(\Lambda(t))-N(\Lambda(s)))} ds \\
&= X(i)2^{-(k+1)} + \int_i^t 2^{-(N(\Lambda(t))-N(\Lambda(s)))} ds
\end{aligned}$$

Therefore, the lemma holds for $N(\Lambda(t)) - N(\Lambda(i)) = k+1$. Since (5.1) holds for $N(\Lambda(t)) - N(\Lambda(i)) = 1$, and since it holds for $N(\Lambda(t)) - N(\Lambda(i)) = k+1$ if it holds for $N(\Lambda(t)) - N(\Lambda(i)) = k$, the result is proven for all integers k. ■

To define the Markov chain, subdivide time into unit intervals, and define $X_i(t) = X(i+t)$. Also define $X_i = \{X_i(t) : 0 \leq t < 1\}$. Elements X_i, which will constitute the Markov chain, reside in the following state space.

Proposition 5.1 *Each element X_i belongs to the space S, where*

$$S = \left\{ x2^{-M(t)} + \int_0^t 2^{-(M(t)-M(s))} ds : t \in [0,1),\ x \in \mathbb{R} \right\}, \tag{5.6}$$

where $M(t)$ is a nonhomogeneous Poisson process on $[0,1)$.

Proof Elements of S are all of the form found in Lemma 5.1. To show that $X_i \in S$,

let $x = X(i) = X_i(0)$. (5.1) can be written as,

$$\begin{aligned} X_i(t) &= X(i+0)2^{-(N(\Lambda(i+t))-N(\Lambda(i+0)))} + \int_{i+0}^{i+t} 2^{-(N(\Lambda(i+t))-N(\Lambda(s)))}ds \\ &= x2^{-M(t)} + \int_0^t 2^{-(N(\Lambda(i+t))-N(\Lambda(i+u)))}du \\ &= x2^{-M(t)} + \int_0^t 2^{-(N(\Lambda(i+t))-N(\Lambda(i+0))-(N(\Lambda(i+u))-N(\Lambda(i+0))))}du \\ &= x2^{-M(t)} + \int_0^t 2^{-(M(t)-M(u))}du, \end{aligned}$$

and $M(t) = N(\Lambda(i+t)) - N(\Lambda(i))$ is a nonhomogeneous Poisson process. ■

Note that elements of S are collections of right-continuous slope-one line segments (see Figure 1). $M(t)$ is the arrival process, on $[0, 1)$, of window reductions due to packet losses, and the points of discontinuity in S correspond to these arrivals. These discontinuities are referred to as jumps in the window size.

One can partition S on the basis of the number of jumps an element has. This will prove to be a useful distinction, since the number of jumps largely determines the behavior of a state.

Definition 5.1 *Let S^n be the set of elements with a total of n jumps. i.e.*

$$S^n = \{x \in S : M(1) = n\}. \tag{5.7}$$

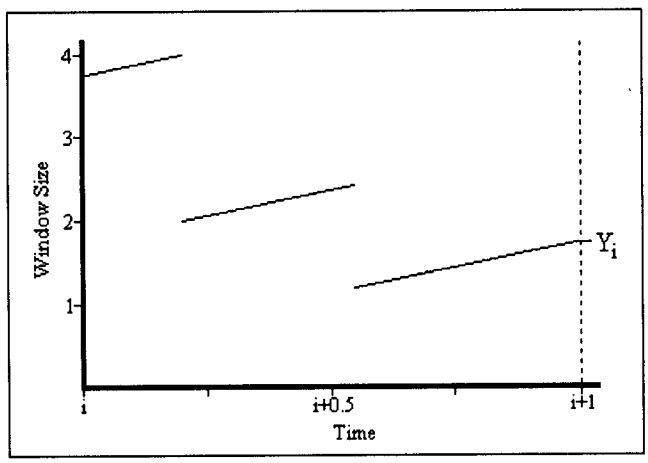

Figure 1: *A typical element X_i in S. The endpoint Y_i is also shown.*

The collections of elements in S which have no jumps is important, and a special subset of this collection receives its own definition:

Definition 5.2 *Let S' be a subset of S (and of S^0) such that*

$$\begin{aligned} S' &= \{x \in S^0 : x(0) \geq 1\} \\ &= \{x \in S^0 : x(1) \geq 2\}. \end{aligned} \tag{5.8}$$

There are technical reasons for defining S' in this way. In the next chapter, the Markov chain of window sizes is followed through a large number of single-jump states followed by a number of zero-jump states. It turns out that, after many single-jump states, $x(0)$ must be larger than one, so S' is the largest zero-jump set that can be reached with positive probability after this type of behavior. If the theory were extended to follow states with more than one jump, then this restriction could be removed. However, such an extension to the theory is neither realistic nor necessary.

We will often need to consider the value at the right-hand endpoint of an interval or of a line segment. Since S is made of increasing, right-continuous line segments, the meaning of a right-hand endpoint must be clarified.

- The value at the right-hand endpoint of an increasing line segment, $l(t)$ is understood to be $\sup_t\{l(t) : t \in [a, b)\}$, where the line segment spans the time interval [a,b).
- The value at the right-hand endpoint of an element $X_i \in S$ is understood to be the value at the right-hand endpoint of the last line segment in X_i.

This convention leads to the following definition.

Definition 5.3 *For any $i \geq 0$, Y_i is a random variable representing the value of the right-hand endpoint of $X_i \in S$, (see Figure 1).*

Using Lemma 5.1 and the state space S, we can now adequately define a Markov chain on a general state space.

Theorem 5.1 *Let $\mathbb{Z}^+ = (0, 1, 2, 3, ...)$. The sequence $\{X_i : i \in \mathbb{Z}^+\}$ constitutes a Markov chain on the state space S.*

Proof To show that $\{X_i\}$ defines a Markov chain, two observations are necessary. First, it can be seen that $X_i(0)$ depends on both Y_{i-1} and $dN(\Lambda(i))$ by considering (5.1) as follows. For any $\varepsilon > 0$,

$$\begin{aligned} X_i(0) &= X(i+0) \\ &= X(i-\varepsilon)2^{-(N(\Lambda(i))-N(\Lambda(i-\varepsilon)))} + \int_{i-\varepsilon}^{i} 2^{-(N(\Lambda(i))-N(\Lambda(s)))}ds \end{aligned}$$

Taking limits, and using the existing right continuity, we obtain

$$\begin{aligned} X_i(0) &= \lim_{\varepsilon \to 0^+} X(i-\varepsilon)2^{-(N(\Lambda(i))-N(\Lambda(i-\varepsilon)))} + \int_{i-\varepsilon}^{i} 2^{-(N(\Lambda(i))-N(\Lambda(s)))}ds \\ &= Y_{i-1}2^{-dN(\Lambda(i))} \end{aligned} \tag{5.9}$$

If $dN(\Lambda(i)) = 0$ (an event with probability 1 for all i), then there are no jumps in $X(t)$ at time i so $X_i(0) = Y_{i-1}$ *a.s.* Furthermore, if there were a jump at time i, then moving the jump to time $i + \delta$ would result in a negligible change to state X_i. Therefore we may take $X_i(0) = Y_{i-1}$ in all that follows.

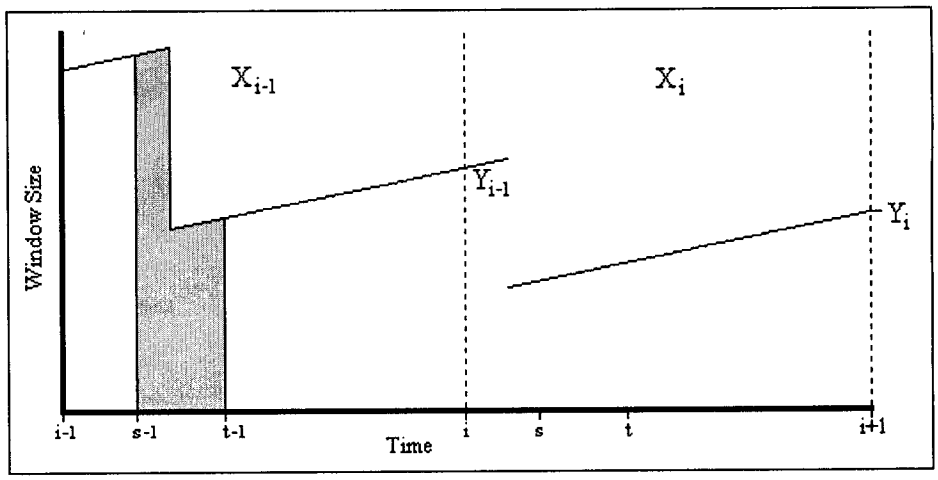

Figure 2: *Two consecutive elements X_{i-1} and X_i. The stochastic intensity of the jump process in the region (s,t) is proportional to the shaded area. Also note $X_i(0) = Y_{i-1}$.*

The second observation concerns the intensity of the Poisson process. This is given by the rate of change of the compensator, which according to (4.14) (the constant loss-probability case), is $\lambda(t) = kX(t-1)$. According to [3], the number of arrivals in a time-varying Poisson process in an interval (s,t) has a Poisson distribution with parameter $\Lambda(s,t)$, where

$$\begin{aligned}\Lambda(s,t) &= \Lambda(t) - \Lambda(s) \\ &= \int_s^t \lambda(u)du \\ &= \int_s^t kX(u-1)du \\ &= \int_{s-1}^{t-1} kX(u)du \end{aligned} \tag{5.10}$$

For any state X_i, the number of jumps in the state is the number of Poisson arrivals between $X(i)$ and $X(i+1)$. This is Poisson distributed with parameter $\int_{i-1}^{i} kX(\tau)d\tau$, and is therefore completely determined by state X_{i-1}. Note that this implies that the value of $dN(\Lambda(i))$ found in (5.9) is also completely determined by state X_{i-1}, in particular by $X_{i-1}(0)$.

These two observations together give the Markov property; for a given loss probability k, state X_i is determined by state X_{i-1}. ■

One final note. The class of Borel sets of S will be denoted by $\mathcal{B}(S)$, which can be constructed in the following manner. For an element of S^n, define an $(n+1)$-dimensional vector in which the first entry represents the left-hand endpoint x of the element, the second entry represents the time until the first jump, the third entry represents the time between the first and second jumps, and so on. An $(n+1)$-dimensional class of Borel sets can be generated out of such elements in the manner described on page 158 of [4]. Such a class contains the elements of S^n as a subset. Taking the countable disjoint union of such classes for all values of n yields a σ-field which contains S as a subset. By Theorem 10.1 of [4], this σ-field can be intersected with S itself to give the appropriate σ-field $\mathcal{B}(S)$.

5.1 Commentary

The Markov chain above was constructed on the space of window size trajectories, under the assumption that the packet loss probability was constant. It is tempting to try to accommodate time-varying loss probabilities by expanding the state space to include loss probability trajectories as well, but the resulting system is non-Markovian; the loss probability depends on a large number of window sizes, all but one of which fall outside the state space. Moreover, attempts to remedy this by creating a state space that includes the window trajectories of all connections fail, since each TCP source evolves on its own local time scale.

Strictly speaking then, all stability properties based on Markov chain theory will only apply to the constant loss-probability case, since this is the only context in which a Markov chain truly occurs. However, the stability arguments of the next chapter still work in the variable loss-probability case, as long as $k(t) \geq k_{min} > 0$ for all t. In this case, (5.10) takes the form,

$$\Lambda(s,t) = \int_{s-1}^{t-1} k(\tau_n(u))X(u)du \tag{5.11}$$

The advantage of proving stability for the variable-loss-probability case is that it demonstrates that window size processes may couple even when the loss probability varies. Since separate copies of the process tend to couple as time goes on, the initial conditions become negligible after a time, and the processes become statistically similar, regardless of initial conditions. Since this is true for time-varying loss probabilities as well as constant loss probabilities, drop policies such as RWFD may still lead to stability in some form.

With this in mind, the next chapter will make stochastic stability arguments assuming a variable nonzero loss probability, although a constant loss probability is implied by the presence of a Markov chain. Constant loss-probability stability, which is the only true stability, may then be considered as a special case.

Chapter 6

Stability of the Markov Chain

If the Markov chain describing the time-changed window size dynamics exhibits some sort of stability, then much can be said (and will be said in the next chapter) about the long-term stability of the TCP bottleneck network under study. The Aperiodic Ergodic Theorem (Theorem 13.0.1 of [2]) can be used to establish stability of a Markov chain in the sense that, in the long run, the state tends to a unique distribution independent of initial conditions. The theorem is paraphrased below.

Theorem 6.1 *(Aperiodic Ergodic Theorem) Suppose that* $\Phi = \{X_n\}$ *is an aperiodic Harris recurrent chain, with invariant measure* π*. If there exists some petite set* C*, some* $b < \infty$ *and a non-negative function* V *finite at some one* $x_0 \in S$*, satisfying*

$$\Delta V(x) = \int_S P(x, dy)V(y) - V(x) \leq -1 + b\chi_C(x), x \in S, \tag{6.1}$$

then there exists a unique invariant probability measure π *(a scaled version of the given* π*) such that for every initial condition* $x \in S$*,*

$$\sup_{A \in \mathcal{B}(S)} |P^n(x, A) - \pi(A)| \to 0 \text{ as } n \to \infty \tag{6.2}$$

$P^n(x, A)$ denotes the n-step transition probability, $P(X_n \in A | X_0 = x)$. The Aperiodic Ergodic Theorem thus implies stability in the sense that the long-run distribution of states tends to some fixed distribution π.

A few notes on the terms used above are necessary. Aperiodicity is the property that the chain can transition from any set in $\mathcal{B}(S)$ with positive measure back to the same set in one step with positive probability. Harris recurrence reflects the property that any set in $\mathcal{B}(S)$ with positive measure is (almost surely) visited by the Markov chain an infinite number of times. An invariant measure π is a measure on $\mathcal{B}(S)$ such that

$$\pi(A) = \int_S \pi(dx)P(x, A), \ A \in \mathcal{B}(S).$$

Finally, to define a petite set, take a general state-space Markov chain and create a sampled version by regarding only states $n_0, n_1, n_2, ...$, where each n_i is chosen randomly according to some probability distribution $a(n)$. A set C is petite if, for any $x \in C$ and any set $B \in \mathcal{B}(S)$, the probability of reaching C from x in one step of the sampled chain is bounded below by some nontrivial measure. Such sets are essential to stability because they allow for the construction of atoms; sets for which the transition probability from every element to a given set is the same. These atoms in turn allow copies of a Markov chain to couple, allowing for stability arguments to be made. All of these concepts and arguments are discussed in great depth in [2], and will be developed or referred to as needed throughout the chapter.

The following proposition establishes a method for satisfying the hypotheses of Theorem 6.1, and gives an outline for the rest of the chapter.

Proposition 6.1 *Suppose the following properties can be established for the Markov chain describing the window size evolution:*

- *The Markov chain is ψ-irreducible.*
- *A petite set C exists, with $C = \{x \in S : x(0) \leq c\}$, for some positive integer c.*
- *The chain is aperiodic.*
- *(6.1) holds for the function $V = x(1)$ on the set C.*

Then the hypotheses of the Aperiodic Ergodic Theorem are satisfied.

ψ-irreducibility is the general state-space analog of irreducibility. It expresses the idea that sets with positive ψ-measure can be reached from anywhere in S with positive probability in a finite amount of time.

Proof ψ-irreducibility establishes a class of communicating states. This provides a context in which to consider aperiodicity and Harris recurrence, since these concepts are generally reserved for irreducible chains.

The existence of a petite set is required so that we can examine the drift criterion, (6.1). If this criterion is satisfied, then not only is the main hypothesis of Theorem 6.1 established, but the chain is also guaranteed to be Harris recurrent with an invariant measure π. To see this, consider Theorems 11.3.4, and 10.4.4 of [2] as follows.

Theorem 11.3.4 states that if the drift criterion is satisfied for a chain, then that chain is positive Harris recurrent. It is then clear from the definitions that the chain is also recurrent. Theorem 10.4.4 then states that such a recurrent chain has a unique invariant measure π. ■

Figure 3 shows the details of how Proposition 6.1 leads to the conclusion of the Aperiodic Ergodic Theorem. ψ-irreducibility will be deduced from irreducibility with respect to a weaker measure φ, which is taken to be ordinary Lebesgue measure placed on the right-hand endpoint of a set, and restricted to elements in S'. The petite set C will be constructed from smaller sets possessing the appropriate properties.

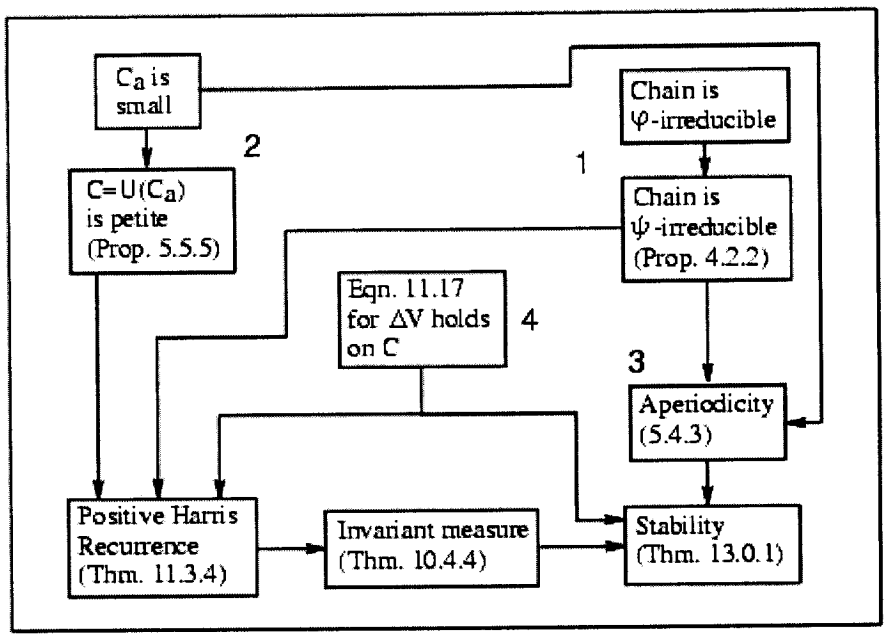

Figure 3: *Satisfying the hypothesis of the Aperiodic Ergodic Theorem. Bold numbers indicate the order in which the four main hypotheses of Proposition 6.1 will be considered.*

The criteria of Proposition 6.1 will be considered, in order, in the sections below.

6.1 ψ-irreducibility of the Chain

On a countable state space, an irreducible Markov chain is one for which any state can be reached from any other with positive probability. Unfortunately, this concept does not directly apply to uncountable state spaces, since the probability of reaching any particular state is generally zero. The analogous concept for Markov chains on a general state-space is the idea of ψ-irreducibility. The idea is that ψ is a measure on the state space such that any set of states with positive ψ measure can eventually be reached from any particular starting state with positive probability.

How do we find ψ? Proposition 4.2.2 of [2] asserts that some maximal irreducibility measure ψ exists if one can find a lesser irreducibility measure φ. Moreover, the

properties of φ also hold for ψ. Consequently, we can look for a simpler irreducibility measure φ with sufficient properties and infer that these properties hold on the maximal measure ψ without having to specify ψ exactly. The candidate for such a measure is specified in the following definition.

Definition 6.1 *For sets in S, let μ represent ordinary Lebesgue measure placed on the right-hand endpoint of a set. That is, $\mu(A)$ is the same as Lebesgue measure on $\{x(1) : x \in A\}$.*

We let φ be a measure on the space S such that,

$$\varphi(A) = \mu(A \cap S'),$$

where S' is defined as in (5.2) to be the set of zero-jump trajectories with left hand endpoint larger than one.

In general, μ will be used to represent both Lebesgue measure on the real line (or a portion thereof), and Lebesgue measure on the endpoint of an interval in S, as in Definition 6.1 above. The exact usage will be clear from the context.

μ itself is not a measure on S, since different elements in S can have the same endpoint. However, restriction of μ to the subset S' does result in a measure. Since elements in S' have no jumps, there is a one-to-one correspondence between such elements and their endpoints. It follows that there is a one-to-one correspondence between the measure of the endpoints of any set $A \cap S'$ and the measure of the set itself. Therefore, φ is a measure since it inherits the required properties from ordinary Lebesgue measure. More precisely, φ is a σ-finite measure with support S'.

To show that a Markov chain is φ-irreducible, the following definition, due to [2] (see Proposition 4.2.1) must be satisfied.

Definition 6.2 *$\Phi = \{X_n\}$ is φ-irreducible if there exists a measure φ on $\mathcal{B}(S)$ such that, for all $x \in S$, whenever $\varphi(A) > 0$, there exists some $n > 0$, possibly depending on both A and x, such that $P^n(x, A) > 0$.*

This definition leads to the main result of this section.

Theorem 6.2 *Suppose the router loss probability is nonzero. Then the Markov chain of window sizes, defined in Theorem 5.1, is φ-irreducible, where φ is as in Definition 6.1.*

Before proving this theorem, a few more definitions and technical lemmas are necessary.

Definition 6.3 *Let $J(X_i)$ be a random variable representing the number of jumps in state X_i. That is,*

$$J(X_i) = N(\Lambda(i+1)) - N(\Lambda(i)) \tag{6.3}$$

Definition 6.4 *For states X_i such that $J(X_i) = 1$, let T_i represent the (local) time of the jump, less 0.5. That is,*

$$T_i = t \in (-0.5, 0.5) \text{ s.t. } dN(\Lambda(i+0.5+t)) = 1. \tag{6.4}$$

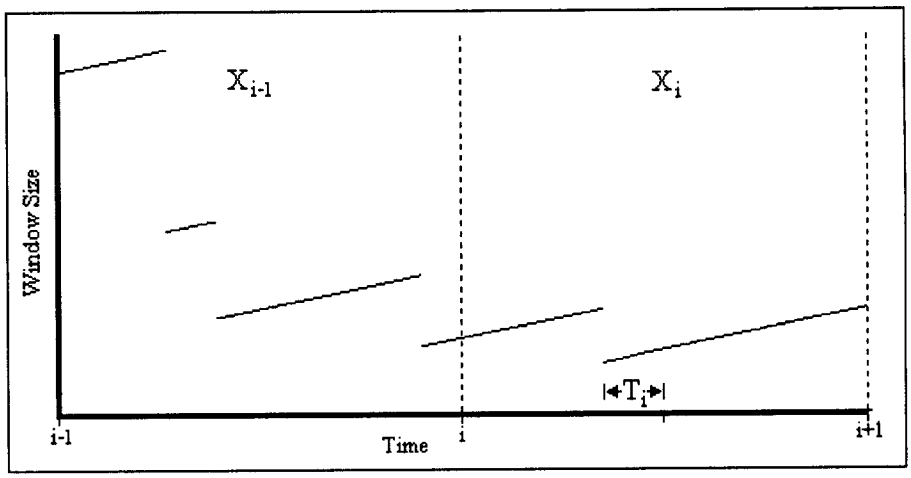

Figure 4: *Two consecutive elements with $J(X_{i-1}) = 3$ and $J(X_i) = 1$. T_i exists and is measured as the distance from the center of the i^{th} interval.*

States with $J(X_i) = 1$ (see Figures 4 and 5) are not uncommon. In fact, they are the primary mechanism by which stability is achieved, since TCP sources will most often cut their window size in response to a single packet loss. To determine the

probability of consecutive single-jump states occuring, we may appeal to a special case of the following lemma.

Lemma 6.1 *Suppose that the loss probability $k(t)$ is nonzero. For any finite integer n, any initial trajectory $X_0 = x$ such that the area below x is finite with $Y_0 = y < \infty$, and any set of integers $\{x_i : 0 \leq x_i < \infty\}$, we have that,*

$$P(J(X_1) = x_1,\ J(X_2) = x_2, ...J(X_n) = x_n | X_0 = x) > 0. \tag{6.5}$$

Proof Starting from y, the window size in each of the n steps must be finite, since it grows at a rate less than or equal to one per step. The stochastic intensity in any of these steps, given by $\Lambda(s,t)$ as in (5.11) is always finite and positive, since $k(t) > 0$. Since the number of jumps in each step has a Poisson distribution with parameter $\Lambda(s,t)$, the outcome of each random variable can take on any allowable value x_i with positive probability. Therefore, the probability of having x_i jumps in any step is always positive.

Finally, note that the number of jumps in any of these steps can take on value x_i with positive probability regardless of the number of jumps in other steps. That is, for any i, and any values of the x_j, $1 \leq j \leq n$,

$$P(J(X_i) = x_i | J(X_{i-1}) = x_{i-1}, ...J(X_1) = x_1) > 0$$

We may now use the definition of conditional probability to write,

$$\begin{aligned} & P(J(X_1) = x_1,\ J(X_2) = x_2, ...J(X_n) = x_n) \\ = \ & P(J(X_n) = x_n | J(X_{n-1}) = x_{n-1}, ...J(X_1) = x_1) \\ & \cdot P(J(X_{n-1}) = x_{n-1} | J(X_{n-2}) = x_{n-2}, ...J(X_1) = x_1) \\ & ... \cdot P(J(X_2) = x_2 | J(X_1) = x_1) P(J(X_1) = x_1) \\ > \ & 0 \end{aligned}$$

Therefore, the intersection of events $\{J(X_i) = x_i\}$ has positive probability. ∎

In light of Lemma 6.1, a series of states in which there is exactly one jump per state X_i (that is, $J(X_i) = 1, 1 \leq i \leq n$) may occur with positive probability. Consider

now such a series of single-jump states. Examining the behavior of the chain when such a series occurs will lead to a few important results, ending with Lemma 6.2.

Since Poisson arrivals can occur at any time, it must be true that T_i can fall in any set of positive Lebesgue measure over (-0.5,0.5) so that,

$$P(T_i \in A) > 0 \; \forall \; A \in \mathcal{B}(\mathbb{R}) \; s.t. \; \mu\left(A \cap (-0.5, 0.5)\right) > 0 \tag{6.6}$$

where μ represents Lebesgue measure.

For these states we have, similar to the proof of Lemma 5.1,

$$N(\Lambda(i+1)) - N(\Lambda(i+t)) = \chi_{\{t<T_i+0.5\}}, \tag{6.7}$$

and as a special case,

$$N(\Lambda(i+1)) - N(\Lambda(i)) = 1. \tag{6.8}$$

Starting with (5.1), we may use (6.7) and (6.8), along with the change of variables $u = s - i$ to write,

$$\begin{aligned}
Y_i = X(i+1) &= X_i(0)2^{-(N(\Lambda(i+1))-N(\Lambda(i)))} + \int_i^{i+1} 2^{-(N(\Lambda(i+1))-N(\Lambda(s)))}ds \\
&= \frac{Y_{i-1}}{2} + \int_i^{i+1} 2^{-(N(\Lambda(i+1))-N(\Lambda(s)))}ds \\
&= \frac{Y_{i-1}}{2} + \int_0^{1} 2^{-(N(\Lambda(i+1))-N(\Lambda(i+u)))}du \\
&= \frac{Y_{i-1}}{2} + \int_0^{T_i+0.5} 2^{-\chi_{\{u\le T_i+0.5\}}}du + \int_{T_i+0.5}^{1} 2^{-\chi_{\{u\le T_i+0.5\}}}du \\
&= \frac{Y_{i-1}}{2} + \int_0^{T_i+0.5} 2^{-1}du + \int_{T_i+0.5}^{1} 2^{-0}du \\
&= \frac{Y_{i-1}}{2} + \left(\frac{T_i}{2} + \frac{1}{4}\right) + \left(1 - T_i - \frac{1}{2}\right) \\
&= \frac{Y_{i-1}}{2} + \frac{3}{4} - \frac{T_i}{2}
\end{aligned} \tag{6.9}$$

The value of Y_i is thus determined by the previous value and by a random time, as in Figure 5.

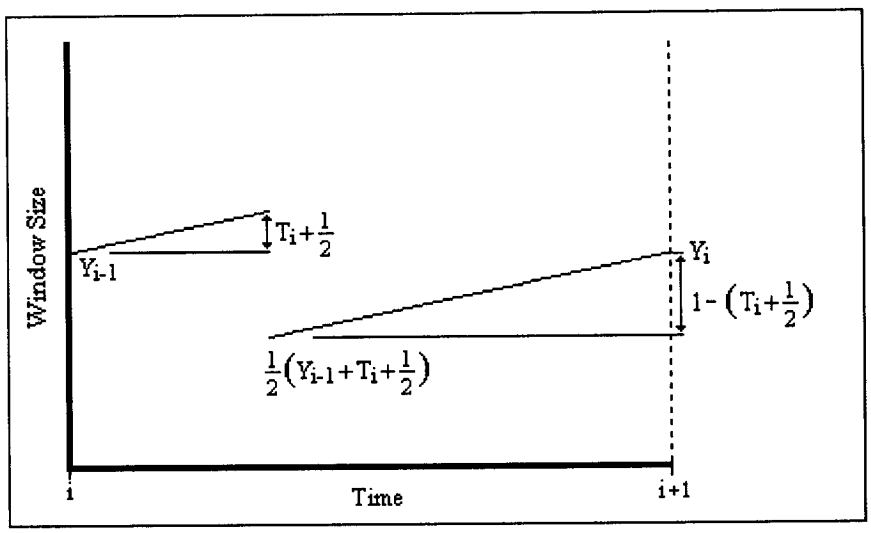

Figure 5: *Y_i as determined by Y_{i-1} and T_i for a single-jump state.*

As time goes on and the number of single-jump transitions becomes large, the contribution of the initial value to the current state becomes negligible. To see this, let $Y_0 = y$. For any finite integer n, we may perform a recursive calculation to obtain

$$\begin{aligned} Y_n &= \frac{Y_{n-1}}{2} + \frac{3}{4} - \frac{T_n}{2} \\ &= \frac{\frac{Y_{n-2}}{2} + \frac{3}{4} - \frac{T_{n-1}}{2}}{2} + \frac{3}{4} - \frac{T_n}{2} \\ &= \frac{Y_{n-2}}{4} + \frac{3}{8} + \frac{3}{4} - \frac{T_{n-1}}{4} - \frac{T_n}{2} \\ &= \cdots \\ &= \frac{y}{2^n} + \frac{3(2^n-1)}{2^{n+1}} - \sum_{i=1}^{n} 2^{-(n-i+1)} T_i \\ &= \frac{3}{2} + \frac{y-3/2}{2^n} - \sum_{i=1}^{n} 2^{-(n-i+1)} T_i \end{aligned} \tag{6.10}$$

The first two terms of (6.10) tend to 3/2 as n approaches infinity. More precisely, for any $\varepsilon > 0$, choose $n \in \mathbb{Z}$ s.t. $n > \log_2((y-3/2)/\varepsilon)$. Then

$$\left|\frac{y-3/2}{2^n}\right| < \varepsilon.$$

The third term of (6.10) is bounded in modulus by a convergent geometric series, since $|T_i| < 1/2$ for all i. That is,

$$\begin{aligned} |\sum_{i=1}^{n} 2^{-(n-i+1)} T_i| &\leq \frac{1}{2} \sum_{i=1}^{n} 2^{-(n-i+1)} \\ &= \frac{1}{2} \sum_{j=1}^{n} 2^{-j} \\ &= \frac{1}{2} - \left(\frac{1}{2}\right)^{n+1} \qquad (6.11) \\ &\to \frac{1}{2} \; as \; n \to \infty. \qquad (6.12) \end{aligned}$$

The value of Y_n is thus determined largely by the recent past, while earlier events have a contribution that decays exponentially. As n gets large, therefore, the initial value of the chain becomes negligible, and the values of Y_n tend to lie in the interval,

$$\begin{aligned} &\left(\frac{3}{2} + \frac{y - 3/2}{2^n} - \left(\frac{1}{2} - \left(\frac{1}{2}\right)^{n+1}\right), \frac{3}{2} + \frac{y - 3/2}{2^n} + \left(\frac{1}{2} - \left(\frac{1}{2}\right)^{n+1}\right)\right) \\ = &\left(1 + \frac{y-1}{2^n}, 2 + \frac{y-2}{2^n}\right) \\ \to &\; (1, 2) \; as \; n \to \infty, \end{aligned}$$

(see Figure 6.)

If Y_n tends to fall in the interval $(1, 2)$ after many single-jump steps, then it is reasonable to expect that it may fall in subsets of this interval which have positive Lebesgue measure. The following lemma formalizes this notion.

Lemma 6.2 *For any finite y such that $Y_0 = y$, and for any subset B of the interval $(1, 2)$ with positive Lebesgue measure, $\exists \; N < \infty$ such that, $\forall \; n \geq N$, if $J(X_i) = 1 \; \forall i \leq n$, then $P(Y_n \in B) > 0$.*

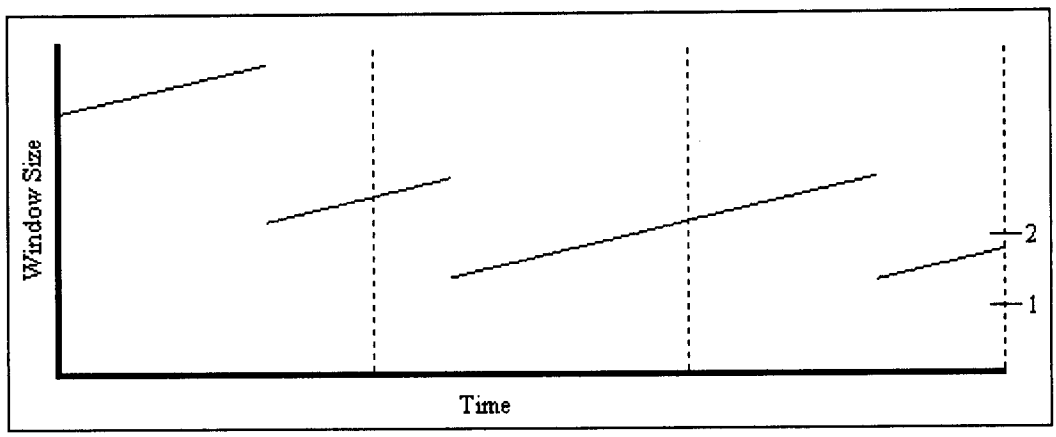

Figure 6: *Starting from any Y_0, the endpoint must always fall in or close to $(1,2)$ after a sufficient number of single-jump steps. The distribution of T_i ensures that it can, with positive probability, fall in any subset of $(1,2)$ which has positive measure.*

Proof The first step in the proof is to choose a target subset of B with positive Lebesgue measure μ. Then, if this target subset can be reached in n steps with positive probability, the result will be proven.

Assume that $\mu(B) = m$, where $m > 0$, and that B itself is contained in the interval $(1,2)$, which may be partitioned into:

$$\begin{aligned} A_1 &= \{x \in \mathbb{R} : 1 < x < 1 + m/4\} \cup \{x \in \mathbb{R} : 2 - m/4 < x < 2\} \\ A_2 &= \{x \in \mathbb{R} : 1 + m/4 \leq x < 3/2\} \\ A_3 &= \{x \in \mathbb{R} : 3/2 \leq x \leq 2 - m/4\} \end{aligned}$$

Note that $\mu(A_1) = m/2$, so it follows that $\mu(B \cap A_1) \leq m/2$. Since,

$$\mu(B) = \mu(B \cap A_1) + \mu(B \cap A_2) + \mu(B \cap A_3),$$

we may establish a lower bound,

$$\mu(B \cap A_2) + \mu(B \cap A_3) \geq m/2.$$

This implies that either, or both, of the following equations must hold:

$$\mu(B \cap A_2) \geq m/4 \tag{6.13}$$

$$\mu(B \cap A_3) \geq m/4. \tag{6.14}$$

Choose either $B \cap A_2$ or $B \cap A_3$ as the target subset of B, depending on which of (6.13) and (6.14) are satisfied. The method for reaching this subset with positive probability will differ, depending on which is chosen.

Regardless of the target subset, the next step in the proof is the same: from the initial point $Y_0 = y$, follow n_1 single-jump steps in the chain in such a way that $1 < Y_{n_1} < 2$. This can be done by requiring that,

$$-1/4 < T_i < 1/4, \ \forall i: \ 1 \leq i \leq n_1.$$

That is, each T_i must fall in an interval with Lebesgue measure 1/2. By (6.6), this may happen with positive probability for each T_i. Indeed it may happen for all the T_i's, since each T_i can fall in the interval with positive probability regardless of the other values. (The argument is similar to that at the end of Lemma 6.1.) Then, by (6.10),

$$\begin{aligned} Y_{n_1} &< \frac{3}{2} + \frac{y - 3/2}{2^{n_1}} + \frac{1}{4}\sum_{i=1}^{n_1} 2^{-(n_1 - i + 1)} \\ &= \frac{7}{4} + \frac{y - 7/4}{2^{n_1}}, \end{aligned}$$

and,

$$\begin{aligned} Y_{n_1} &> \frac{3}{2} + \frac{y - 3/2}{2^{n_1}} - \frac{1}{4}\sum_{i=1}^{n_1} 2^{-(n_1 - i + 1)} \\ &= \frac{5}{4} + \frac{y - 5/4}{2^{n_1}}. \end{aligned}$$

Now simply choose n_1 large enough so that,

$$\left| \frac{y - 7/4}{2^{n_1}} \right| < \frac{1}{4},$$

and,

$$\left| \frac{y - 5/4}{2^{n_1}} \right| < \frac{1}{4}.$$

This will guarantee that $1 < Y_{n_1} < 2$. This first step establishes a general starting point. The next step in the proof takes the chain to a position from which it can hit the target set in one step of the chain.

For the case that the target subset of B is $B \cap A_2$, Let

$$1/2 - \varepsilon < T_i < 1/2, \text{ for all } i: \ n_1 + 1 \leq i \leq n_1 + n_2,$$

where n_2 and $\varepsilon > 0$ are chosen so that,

$$\begin{aligned} n_2 &> \log_2\left(\frac{2(Y_{n_1} - 1)}{m}\right), \\ \varepsilon &< \frac{m/2 + (1 - Y_{n_1})/2^{n_2}}{1 - 1/2^{n_2}}. \end{aligned}$$

The choice of n_2 guarantees that an $\varepsilon > 0$ may be found. Therefore, the T_i may lie in the prescribed sets with positive probability, since each set has positive Lebesgue measure ε. Again, by (6.10),

$$\begin{aligned} Y_{n_1+n_2} &< \frac{3}{2} + \frac{Y_{n_1} - 3/2}{2^{n_2}} - \left(\frac{1}{2} - \varepsilon\right) \sum_{i=1}^{n_2} 2^{-(n_2 - i + 1)} \\ &= \frac{3}{2} + \frac{Y_{n_1} - 3/2}{2^{n_2}} - \left(\frac{1}{2} - \varepsilon\right)\left(1 - \frac{1}{2^{n_2}}\right) \\ &= 1 + \frac{Y_{n_1} - 1}{2^{n_2}} + \varepsilon\left(1 - \frac{1}{2^{n_2}}\right) \\ &< 1 + m/2, \end{aligned}$$

by the choice of ε. Also,

$$\begin{aligned} Y_{n_1+n_2} &> \frac{3}{2} + \frac{Y_{n_1} - 3/2}{2^{n_2}} - \frac{1}{2} \sum_{i=1}^{n_2} 2^{-(n_2 - i + 1)} \\ &= \frac{3}{2} + \frac{Y_{n_1} - 3/2}{2^{n_2}} - \frac{1}{2}\left(1 - \frac{1}{2^{n_2}}\right) \\ &= 1 + \frac{Y_{n_1} - 1}{2^{n_2}} \\ &> 1, \end{aligned}$$

since $Y_{n_1} > 1$.

For the T_i lying in the appropriate sets as above, it is therefore guaranteed that $1 < Y_{n_1+n_2} < 1 + m/2$. Moreover, the discussion above, based on Lemma 6.1 and (6.6), establishes that the event

$$\{J(X_i) = 1 \ \forall \ i \leq n, \ |T_i| < 1/4 \ \forall \ i \leq n_1, \ 1/2 - \varepsilon < T_i < 1/2 \ \forall \ n_1 < i \leq n_2\}$$

has positive probability.

It can now be shown that, starting from such a point $Y_{n_1+n_2}$, the target subset $B \cap A_2$ may be reached with positive probability in a single step, so that $n = n_1+n_2+1$. We write the whole calculation as,

$$\begin{aligned} P^n(y,B) &= P(Y_n \in B | Y_0 = y) \\ &\geq P(Y_n \in B \cap A_2 | Y_0 = y, J(X_i) = 1 \ \forall \ i \leq n, \\ &\quad |T_i| < 1/4 \ \forall \ i \leq n_1, \ 1/2 - \varepsilon < T_i < 1/2 \ \forall \ n_1 < i \leq n_2) \\ &\quad \cdot P(J(X_i) = 1 \ \forall \ i \leq n, \ |T_i| < 1/4 \ \forall \ i \leq n_1, \\ &\quad 1/2 - \varepsilon < T_i < 1/2 \ \forall \ n_1 < i \leq n_2) \\ &= P(Y_n \in B \cap A_2 | 1 < Y_{n-1} < 1 + m/2, ...) \cdot P(...) \end{aligned} \tag{6.15}$$

We may now transform this into a probability calculation for T_n. From (6.9),

$$\begin{aligned} Y_n &= \frac{Y_{n-1}}{2} + \frac{3}{4} - \frac{T_n}{2} \\ &= \frac{Y_{n_1+n_2}}{2} + \frac{3}{4} - \frac{T_n}{2}. \end{aligned} \tag{6.16}$$

Let $Z = Y_{n-1}/2 + 3/4$. Then the event $\{1 < Y_{n-1} < 1 + m/2\}$ is equivalent to

$$\left\{ \frac{5}{4} < Z < \frac{5}{4} + \frac{m}{4} \right\},$$

and (6.16) is transformed into

$$Y_n = Z - \frac{T_n}{2}. \tag{6.17}$$

Let $\psi_y(B)$ be a set mapping such that $\{b \in B\} \rightarrow \{-2b + 2y\}$. (Then $\psi_Z(\{Z - T_n/2\}) = \{T_n\}$.) Note that ψ_y is an invertible mapping for fixed y, so $x \in B$ if and only if $\psi_y(x) \in \psi_y(B)$. Applying ψ_Z to (6.15), we may write

$$\begin{aligned} P^n(y,B) &\geq P(\psi_Z(\{Z - T_n/2\}) \in \psi_Z(B \cap A_2) | 1 < Y_{n-1} < 1 + m/2, ...) \cdot P(...) \\ &= P\left(T_n \in \psi_Z(B \cap A_2) \middle| \frac{5}{4} < Z < \frac{5}{4} + \frac{m}{4}, ... \right) \cdot P(...). \end{aligned} \tag{6.18}$$

Note that $\mu(\psi_Z(B \cap A_2)) > 0$, since ψ_y only doubles the Lebesgue measure of a set. Calculating the minimum and maximum values of the mapping $\psi_Z(B \cap A_2)$, for $5/4 < Z < 5/4 + m/4$, and $B \cap A_2 \subset (1 + m/4, 3/2))$, we find,

$$\begin{aligned} \min(\psi_Z(B \cap A_2)) &= -2\max(b : b \in B \cap A_2) + 2\min(y : 5/4 < y < 5/4 + m/4) \\ &= -2\left(\frac{3}{2}\right) + 2\left(\frac{5}{4}\right) \\ &= -0.5, \end{aligned}$$

and,

$$\begin{aligned} \max(\psi_Z(B \cap A_2)) &= -2\min(b : b \in B \cap A_2) + 2\max(y : 5/4 < y < 5/4 + m/4) \\ &= -2\left(\frac{4+m}{4}\right) + 2\left(\frac{5+m}{4}\right) \\ &= 0.5. \end{aligned}$$

Therefore, $\psi_Z(B \cap A_2) \subset (-0.5, 0.5)$. Since $\psi_Z(B \cap A_2)$ has positive Lebesgue measure and is contained in the interval $(-0.5, 0.5)$, T_n may fall in this set with positive probability by (6.6). Therefore, $P^n(y, B) > 0$, for large enough n.

The proof is similar if the target subset is $B \cap A_3$, and only an abbreviated version is given here. Let

$$-1/2 < T_i < -1/2 + \varepsilon, \text{ for all } i : \ n_1 + 1 \leq i \leq n_1 + n_2,$$

where n_2 and $\varepsilon > 0$ are chosen so that,

$$\begin{aligned} n_2 &> \log_2\left(\frac{2(2 - Y_{n_1})}{m}\right), \\ \varepsilon &< \frac{m/2 - (2 - Y_{n_1})/2^{n_2}}{1 - 1/2^{n_2}}. \end{aligned}$$

The choice of n_2 guarantees that an $\varepsilon > 0$ may be found. By (6.10), this choice of ε gives

$$2 - m/2 < Y_{n_1+n_2} < 2.$$

Moreover, the event

$$\{J(X_i) = 1 \ \forall \ i \leq n, \ |T_i| < 1/4 \ \forall \ i \leq n_1, \ -1/2 < T_i < -1/2 + \varepsilon \ \forall \ n_1 < i \leq n_2\}$$

has positive probability as before.

Starting from such a point $Y_{n_1+n_2}$, the target subset $B \cap A_3$ may be reached with positive probability in a single step, with $n = n_1 + n_2 + 1$. We may condition on the above event to write,

$$\begin{aligned} P^n(y,B) &= P(Y_n \in B | Y_0 = y) \\ &\geq P(Y_n \in B \cap A_3 | Y_0 = y, J(X_i) = 1 \ \forall \ i \leq n, \\ &\quad -1/4 < T_i < 1/4 \ \forall \ i \leq n_1, \ -1/2 < T_i < -1/2 + \varepsilon \ \forall \ n_1 < i \leq n_2) \\ &\quad \cdot P(J(X_i) = 1 \ \forall \ i \leq n, \ -1/4 < T_i < 1/4 \ \forall \ i \leq n_1, \\ &\quad -1/2 < T_i < -1/2 + \varepsilon \ \forall \ n_1 < i \leq n_2) \\ &= P(Y_n \in B \cap A_3 | 2 - m/2 < Y_{n-1} < 2, ...) \cdot P(...) \end{aligned} \tag{6.19}$$

For $Z = Y_{n-1}/2 + 3/4$, the event $\{2 - m/2 < Y_{n-1} < 2\}$ is equivalent to

$$\left\{ \frac{7}{4} - \frac{m}{4} < Z < \frac{7}{4} \right\}.$$

Let $\psi_y(B)$ be a set mapping as before, and apply ψ_Z to (6.19) to obtain

$$\begin{aligned} P^n(y,B) &\geq P(\psi_Z(\{Z - T_n/2\}) \in \psi_Z(B \cap A_3) | 2 - m/2 < Y_{n-1} < 2, ...) \cdot P(...) \\ &= P\left(T_n \in \psi_Z(B \cap A_3) \middle| \frac{7}{4} - \frac{m}{4} < Z < \frac{7}{4}, ... \right) \cdot P(...). \end{aligned} \tag{6.20}$$

Note that $\mu(\psi_Z(B \cap A_3)) > 0$. Calculating the minimum and maximum values of the mapping $\psi_Z(B \cap A_3)$, for the allowable range of Z and $B \cap A_3$, we again find that $\psi_Z(B \cap A_3) \subset (-0.5, 0.5)$. By (6.6), T_n may fall in this set with positive probability. Therefore, $P^n(y,B) > 0$, for large enough n. ∎

We are now ready to prove the main theorem of this section.

Proof of Theorem 6.2. We must show that $P^n(x,A) > 0$ for some n whenever $\varphi(A) > 0$. For the measure φ, it is true that,

$$\varphi(A) > 0 \Rightarrow \exists \ A' \subset A \ s.t. \ \varphi(A') > 0. \tag{6.21}$$

From the definition of φ, this A' can be chosen such that,

$$A' \subset \{x \in S^0 : k \leq x(0) \leq k+1\}$$

for some $k \in \mathbb{Z}^+, 1 \leq k < \infty$. A can be reached in n steps if the subset A' can. For this same k, let

$$A'' = \{x \in S^0 : x + (k-1) \in A'\},$$

so that A'' is a shifted version of A' and,

$$A'' \subset \{x \in S^0 : 1 \leq x(0) \leq 2\}. \tag{6.22}$$

Note that $\varphi(A'') = \varphi(A') > 0$. Also note that since $A'' \subset S'$,

$$\varphi(A'') = \mu(A'' \cap S') = \mu(A'') > 0. \tag{6.23}$$

We will show that the chain may enter A with positive probability by first entering A'' via single-jump states, and then reaching A' via zero-jump states. (See Figure 7.)

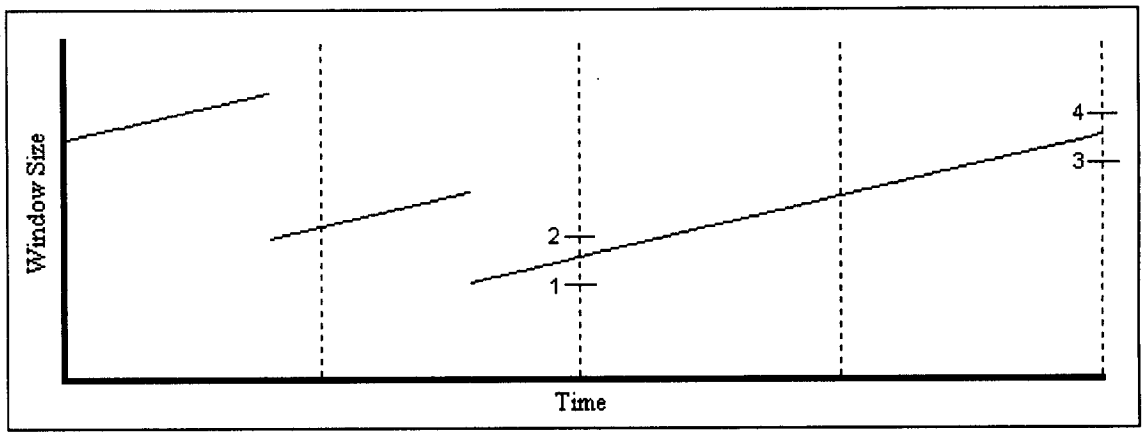

Figure 7: *Reaching A' via A'' via a series of single-jump and zero-jump states.*

Take any finite element $x \in S$ and let the right-hand endpoint of x be denoted y. Let x be the first element of a chain, $x = X_0$. Then $y = Y_0 < \infty$. By Lemma 6.2, there is then a positive probability of reaching any subset of the interval $(1, 2)$ (with

positive Lebesgue measure) in a finite number of single-jump steps. Therefore, there is a positive probability of reaching the set A''.

To reach the set A' from A'', it is merely necessary that there be no jumps in $X_{n+1}, ..., X_{n+k-1}$. By Lemma 6.1, the probability of this event for a finite k is positive. Therefore, there is a positive probability of reaching A' from any element x in a finite number of steps. Since $A' \subset A$, the result is proven.

There is a minor difficulty if the initial state x jitters just above zero for the entire interval. This is the case where the area under x is zero. This gives a jump intensity of zero over the next interval, so that the proof above is invalid. However, this difficulty may be resolved by observing that, in the next step, no jumps will occur and a new state with a nonzero integral will be reached. From here, A can be reached in exactly the same manner as above, so the result still holds with one additional step. ■

We have shown that the chain is φ-irreducible. That is, starting from any state x in the state space, it is possible to reach a set A if $\varphi(A) > 0$. Sets with positive φ-measure therefore provide common ground for the state space, making it impossible to separate S into smaller pieces. This is the major implication of irreducibility; that the dynamics under study may indeed be described using no more than one Markov chain.

Theorem 4.0.1 of [2] infers the existence of a maximal irreducibility measure ψ from that of φ. The exact measure ψ is unknown, but in addition to the properties of φ, it can be said that,

$$\text{if } \psi(A) = 0, \text{ then } \psi\{y : L(y, A) > 0\} = 0.$$

$L(y, A)$ is the probability that the chain can ever reach A starting from y. This is certainly a much stronger property than φ-irreducibility endows. However, φ-irreducibility offers sufficient practical structure for our purposes. The theoretical implications of ψ are suppressed here, and the reader is directed to [2] for further material.

6.2 The Petite Set C

The next step in the verification of Theorem 6.1 is to find a petite set within the state space S. Petite sets are a key ingredient of stochastic stability, as discussed in [2], and the definitions of small and petite sets ([2] pg. 106, 121) must be given before proceeding.

Definition 6.5 *A set $C \in \mathcal{B}(S)$ is called a* small set *if there exists an $m > 0$, and a non-trivial measure ν_m on $\mathcal{B}(S)$, such that for all $x \in C, B \in \mathcal{B}(S)$,*

$$P^m(x, B) \geq \nu_m(B).$$

In the following definition, $a = \{a(n),\ n \geq 1\}$ is a probability distribution on $\mathbb{Z}^+$, and K_a is a probability transition kernel of the sampled chain defined by

$$K_a(x, A) = \sum_{n=0}^{\infty} P^n(x, A)a(n),\ x \in S,\ A \in \mathcal{B}(S).$$

Definition 6.6 *A set $C \in \mathcal{B}(S)$ is called* ν_a-petite *if the sampled chain satisfies the bound*

$$K_a(x, B) \geq \nu_a(B),$$

for all $x \in C, B \in \mathcal{B}(S)$, where ν_a is a non-trivial measure on $\mathcal{B}(S)$.

Theorem 5.2.2 of [2] guarantees the existence of small sets for ψ-irreducible chains, but gives no information on their structure. However, information on the structure of these sets is necessary for the purposes of stability, so specific small and petite sets are constructed in this section.

A small set is, trivially, a petite set, and a finite union of petite sets is also petite (see [2] pp. 121-122). This is a useful observation which will be used in the proof of Theorem 6.3, which builds a petite set from a finite number of small ones. These sets take the following forms.

Let c be a finite positive integer such that $(c+1)(\text{mod } 3) = 0$ (i.e. $c = 2, 5, 8, 11, ...$). Define, for $a \in \mathbb{Z}^+$,

$$C_a = \left\{ x \in S : x(0) < c, \frac{3a}{4} \leq x(1) < \frac{3(a+1)}{4} \right\} \tag{6.24}$$

Also define

$$C = \bigcup_{a=0}^{\frac{4c}{3}+\frac{1}{3}} C_a. \tag{6.25}$$

c is chosen as above so that the upper limit of the union is an integer. For example, if $c = 8$, then the last set in the union is C_{11}, which has as a least upper bound the element $x : x(0) = 8,\ x(1) = 9$, or, $x = 8 + t,\ 0 \leq t \leq 1$. By its construction, C is the set of all trajectories in S which fall entirely below the line $x = c + t,\ 0 \leq t \leq 1$ (see Figure 8).

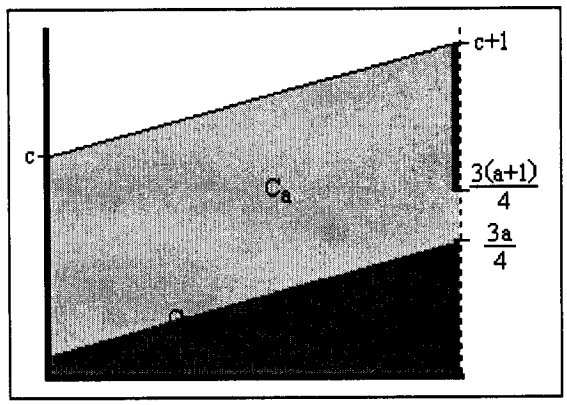

Figure 8: *A small set C_a and the petite set C.*

The main results of this section are given below.

Theorem 6.3 *Assume that the router loss probability $k(t)$ is greater than some nonzero k_{min} for all values of t. Then for any $a \geq 1$, C_a is small with measure*

$$\nu_3^a(B) = \alpha \cdot \frac{6k_{min}}{5}\mu(B \cap Q_a \cap S^0),$$

for some $\alpha > 0$. Here, μ is the Lebesgue measure on the right-hand endpoint, and

$$Q_a = \left\{ x \in S : \frac{3a+11}{8} \leq x(0) < \frac{3a+12}{8} \right\}.$$

To ensure that Theorem 6.3 fits the definitions in (6.24) and (6.25), take $c \geq (3a+7)/4$ in the set C_a. Note that, by the linear increase of elements in S^0, $Q_a \cap S^0$ can be

expressed as,

$$Q_a \cap S^0 = \left\{ x \in S^0 : \frac{3a+19}{8} \leq x(1) < \frac{3a+20}{8} \right\}.$$

This theorem leads directly to the following Corollary.

Corollary 6.1 *Assume that the router loss probability $k(t)$ is greater than some nonzero k_{min} for all t. Then the sets C_0 and C are petite.*

The rest of this section is dedicated to the proofs of Theorem 6.3 and Corollary 6.1. To prove Theorem 6.3, we will focus on what happens to a connection's window size when it starts in a small set C_a and goes through three steps of the chain. These three states will alternate between containing no jumps and single jumps, (see Figure 9.) The following lemma, applied to the first and second steps in this chain, will be essential to the proof.

Lemma 6.3 *Suppose that we are given the following information on two consecutive states in the Markov chain:*

- $J(X_{i-1}) = 0$,
- $J(X_i) = 1$,
- $X_{i-1}(1) = X_i(0) \in A$,
- $A \in \mathbb{R}$ *contains no element smaller than* ρ *for some* $\rho > 1$.

Further suppose that the loss probability $k(t) \geq k_{min}$ for all t. Let the above set of conditions be denoted ξ. If T_i represents the time of the jump (less 0.5) in state X_i, then for any Borel set $D \subset (-0.5, 0.5)$, the probability that the jump time lands in D is bounded below by,

$$P(T_i \in D|\xi) \geq \frac{k_{min}}{1 + \frac{1}{2(\rho-1)}} \mu(D),$$

where μ is Lebesgue measure on $\mathcal{B}(\mathbb{R})$.

Note that the condition $J(X_i) = 1$ requires that k_{min} be nonzero.

Proof Given that $J(X_i) = 1$, T_i represents the conditional time of a single Poisson arrival in a one-second time interval. The Poisson process in question has a time-varying rate which is proportional to the value of $X(t)$ one time unit in the past.

It is well-established that, for a time-varying Poisson process, an arrival time T follows the equation,

$$P(T \leq s | N(1) = 1) = \frac{\Lambda(s)}{\Lambda(1)}. \tag{6.26}$$

(See [3], for example.) In the present formulation, this translates into,

$$P(T_i + 0.5 \leq s | J(X_i) = 1) = \frac{\Lambda(i, i+s)}{\Lambda(i, i+1)}. \tag{6.27}$$

From (5.11) it is known that,

$$\begin{aligned} \Lambda(i, i+s) &= \int_{i-1}^{i+s-1} k(\tau_n(u)) X(u) du \\ &= \int_0^s k(\tau_n(i+u-1)) X_{i-1}(u) du \end{aligned}$$

Since X_{i-1} has no jumps, it is a line segment of the form

$$X_{i-1}(t) = (X_{i-1}(1) - 1) + t, \ 0 \leq t \leq 1.$$

Therefore,

$$\begin{aligned} \Lambda(i, i+s) &= \int_0^s k(\tau_n(i+u-1))((X_{i-1}(1) - 1) + u) du \\ &= \int_0^s k(\tau_n(i+u-1))(\omega + u) du, \end{aligned} \tag{6.28}$$

where $\omega = X_{i-1}(1) - 1 \geq \rho - 1 > 0$.

Putting (6.27) and (6.28) together, and noting that ξ encompasses the appropriate events, we obtain the equation,

$$\begin{aligned} P(T_i + 0.5 \leq s | \xi) &= \frac{\int_0^s k(\tau_n(i+u-1))(\omega + u) du}{\int_0^1 k(\tau_n(i+u-1))(\omega + u) du} \\ &\geq \frac{\int_0^s k_{min}(\omega + u) du}{\int_0^1 (\omega + u) du} \\ &= \frac{k_{min} \int_0^s (\omega + u) du}{\omega + 0.5} \qquad (6.29) \\ &\geq \frac{k_{min} \omega s}{\omega + 0.5}. \qquad (6.30) \end{aligned}$$

This can easily be re-written for Borel sets. Let $D \subset (0,1)$ be a Borel set with corresponding Lebesgue measure $\mu_{(0,1)}$. Then,

$$\begin{aligned} P(T_i + 0.5 \in D|\xi) &= \frac{\int_D k(\tau_n(i+u-1))(\omega+u)\mu_{(0,1)}(du)}{\int_0^1 k(\tau_n(i+u-1))(\omega+u)du} \\ &\geq \frac{k_{min}\int_D(\omega+u)\mu_{(0,1)}(du)}{\omega+0.5} \\ &\geq \frac{k_{min}}{\omega+0.5}\int_D \omega\mu_{(0,1)}(du) \\ &= \frac{k_{min}\omega}{\omega+0.5}\mu_{(0,1)}(D) \\ &= \frac{k_{min}}{1+\frac{1}{2\omega}}\mu_{(0,1)}(D) \end{aligned} \tag{6.31}$$

Here, $D \subset (0,1)$, but the result holds for $D \subset (-0.5, 0.5)$ (and Lebesgue measure $\mu_{(-0.5,0.5)}$ on the interval (-0.5,0.5)) in the form:

$$P(T_i \in D|\xi) \geq \frac{k_{min}}{1+\frac{1}{2\omega}}\mu_{(-0.5,0.5)}(D) \tag{6.32}$$

Since $\omega \geq \rho - 1$, the result is proven. The measure $\mu_{(-0.5,0.5)}(D)$ can be replaced with Lebesgue measure $\mu(D)$ over the whole interval, since D is already restricted to $(-0.5, 0.5)$. ■

The key simplifying feature of this bound is that it depends only on the Lebesgue measure of the set D, which is invariant under translation (as long as D still falls inside the interval $(-0.5, 0.5)$).

We are now ready to prove that C_a is small for $a \geq 1$.

Proof of Theorem 6.3

Let $X_0 = x_0 \in C_a$ for some $a \in \mathbb{Z}^+$, $a \geq 1$, and an appropriate c. Let J_3 be the event,

$$J_3 = \{J(X_1) = 0, J(X_2) = 1, J(X_3) = 0\}.$$

We will associate the probability of such an event with α, so that

$$P(J_3|X_0 \in C_a) \geq \alpha.$$

Since $X_0 \in C_a$, with $X_0(0) < c < \infty$, a simple application of Lemma 6.1 shows that the event J_3 has positive probability. Hence $\alpha > 0$. Note that α is bounded away from zero, since $\Lambda(2,3)$ is nonzero (refer to Figure 9).

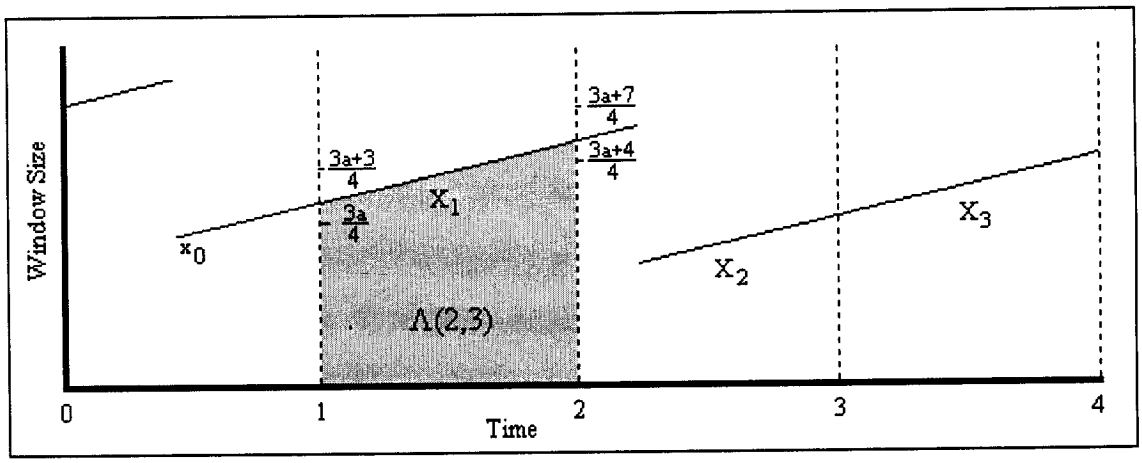

Figure 9: *The event J_3 for $X_0 \in C_a$. $\Lambda(2,3)$ is shaded and must be nonzero, so the required jumps can occur.*

Although the probability of the event J_3 varies with the values of $x_0 \in C_a$ and with T_2, denote the minimum probability over all such possible values by α so that $\alpha > 0$ is fixed.

To show that C_a is small, let ν_3^a be a measure on $\mathcal{B}(S)$ such that,

$$\nu_3^a(\cdot) = \alpha\beta\mu(\cdot \cap Q_a \cap S^0).$$

Here,

$$\beta = \frac{2k_{min}}{1+\frac{2}{3}} = \frac{6k_{min}}{5} > 0,$$

μ is the Lebesgue measure on the right-hand endpoint, and

$$Q_a = \left\{ x \in S : \frac{3a+11}{8} \leq x(0) < \frac{3a+12}{8} \right\}.$$

Since $\alpha > 0$, ν_3^a is indeed a measure. It will be shown that, $\forall\ x_0 \in C_a,\ B \in \mathcal{B}(S)$, we have $P^3(x_0, B) \geq \nu_3^a(B)$, so that C_a is 3-small.

The first step of the proof is to calculate the 3-step transition probabilities conditional on the event J_3. Take $x_0 \in C_a$. The transition kernel defines a regular conditional probability, so

$$\begin{aligned} P^3(x_0, B) &= P(X_3 \in B | X_0 = x_0) \\ &\geq P(X_3 \in B | X_0 = x_0, J_3) \cdot P(J_3) \\ &\geq \alpha P(X_3 \in B | X_0 = x_0, J_3). \end{aligned} \tag{6.33}$$

Consider the possible values of X_1 under the conditions $X_0 = x_0,\ J_3$, where $x_0 \in C_a$ (refer to Figure 9). J_3 implies $J(X_1) = 0$, so $X_1 \in S^0$ and is simply a slope-one line starting at $x_0(1)$. That is,

$$X_1 = x_0(1) + t,\ 0 \leq t \leq 1\ a.s.$$

Since $x_0 \in C_a$, we have

$$\frac{3a}{4} \leq x_0(1) \leq \frac{3(a+1)}{4}.$$

So with probability 1,

$$X_1 \in \left\{ x \in S^0 : \frac{3a}{4} \leq x(0) < \frac{3(a+1)}{4} \right\}.$$

Since X_1 has no jumps, we can make an equivalent restriction on the right-hand endpoints:

$$\begin{aligned} X_1 \quad &\in \quad \left\{ x \in S^0 : \frac{3a}{4} + 1 \leq x(1) < \frac{3(a+1)}{4} + 1 \right\} \\ &= \quad \left\{ x \in S^0 : \frac{3a+4}{4} \leq x(1) < \frac{3a+7}{4} \right\}. \end{aligned} \tag{6.34}$$

Denote this last set of trajectories by A. We have established that, given $X_0 = x_0 \in C_a$ and J_3, we have $X_1 \in A\ a.s.$ Conversely,

$$\forall\ D \subset A^c,\ P(X_1 \in D | X_0 = x_0,\ J_3) = 0. \tag{6.35}$$

Conditioning on the value of X_1 and employing (6.35), (6.33) now gives

$$\begin{aligned} & P^3(x_0, B) \\ \geq\ & \alpha \int_S P(X_3 \in B | X_1 = x_1, X_0 = x_0, J_3) P(X_1 = dx_1 | X_0 = x_0, J_3) \\ =\ & \alpha \int_A P(X_3 \in B | X_1 = x_1, X_0 = x_0, J_3) P(X_1 = dx_1 | X_0 = x_0, J_3) \end{aligned} \tag{6.36}$$

If we can find a lower bound ζ for $P(X_3 \in B|X_1 = dx_1, X_0 = x_0, J_3)$ over all $dx_1 \in A$, then (6.36) will give, for $x_0 \in C_a$,

$$\begin{aligned} P^3(x_0, B) &\geq \alpha \int_A \zeta P(X_1 = dx_1|X_0 = x_0, J_3) \\ &= \alpha\zeta. \end{aligned} \tag{6.37}$$

Clearly, the goal is to find $\zeta = \beta \cdot \mu(B \cap Q_a \cap S^0)$. To do so, we will restrict our attention to a subset of B and condition on the value of X_2 as follows. For all $dx_1 \in A$,

$$\begin{aligned} & P(X_3 \in B|X_1 = dx_1, X_0 = x_0, J_3) \\ \geq\ & P(X_3 \in B \cap Q_a \cap S^0|X_1 = x_1, X_0 = x_0, J_3) \\ =\ & \int_S P(X_3 \in B \cap Q_a \cap S^0|X_2 = x_2, X_1 = x_1, X_0 = x_0, J_3) \\ & \cdot P(X_2 = dx_2|X_1 = x_1, X_0 = x_0, J_3) \end{aligned} \tag{6.38}$$

But, given J_3, we know that $J(X_3) = 0$ and hence $X_3 = X_2(1)+t$, $0 \leq t \leq 1$. Thus X_3 is completely determined by X_2 and J_3, so the first probability in the integrand is either zero or one. Since $X_3 \in S^0$, we know that $X_3 \in B \cap Q_a \cap S^0$ if and only if $X_3(0) \in (B \cap Q_a \cap S^0)(0)$, or equivalently, if $X_2(1) \in (B \cap Q_a \cap S^0)(0)$. (See Figure 9, the notation $A(0)$ refers to the left-hand set of endpoints of a set of trajectories A, and the notation $A(1)$ refers to the right-hand set of endpoints.)

Therefore,

$$\begin{aligned} & P(X_3 \in B \cap Q_a \cap S^0|X_2 = x_2, X_1 = x_1, X_0 = x_0, J_3) \\ =\ & \chi\{x_2(1) \in (B \cap Q_a \cap S^0)(0)\} \end{aligned} \tag{6.39}$$

Substituting (6.39) into (6.38) gives,

$$\begin{aligned} & P(X_3 \in B|X_1 = x_1, X_0 = x_0, J_3) \\ \geq\ & \int_S \chi\{x_2(1) \in (B \cap Q_a \cap S^0)(0)\} P(X_2 = dx_2|X_1 = x_1, X_0 = x_0, J_3) \\ =\ & \int_{\{x_2 : x_2(1) \in (B \cap Q_a \cap S^0)(0)\}} P(X_2 = dx_2|X_1 = x_1, X_0 = x_0, J_3) \\ =\ & P(X_2(1) \in (B \cap Q_a \cap S^0)(0)|X_1 = x_1, X_0 = x_0, J_3) \\ =\ & P(Y_2 \in (B \cap Q_a \cap S^0)(0)|X_1 = x_1, X_0 = x_0, J_3) \end{aligned} \tag{6.40}$$

Now, recall that by (6.34),

$$X_1 \in A = \left\{ x \in S^0 : \frac{3a+4}{4} \leq x(1) < \frac{3a+7}{4} \right\},$$

which implies that,

$$Y_1 \in A_1 = \left\{ y \in \mathbb{R} : \frac{3a+4}{4} \leq y < \frac{3a+7}{4} \right\}. \tag{6.41}$$

That is, the information $Y_1 \in A_1$ is already present by the assumption that $X_1 \in A$, so we can make it explicit without changing anything. (6.40) may therefore be written as,

$$\begin{aligned} & P(X_3 \in B | X_1 = x_1, X_0 = x_0, J_3) \\ \geq\ & P(Y_2 \in (B \cap Q_a \cap S^0)(0) | Y_1 \in A_1, X_1 = x_1, X_0 = x_0, J_3). \end{aligned} \tag{6.42}$$

(6.42) can now be calculated by transforming it into a probability calculation on the local jump time T_2, since T_2 heavily influences the value of Y_2.

Since X_2 has only one jump, we can use (6.9) to relate Y_2 to Y_1 by

$$Y_2 = \frac{Y_1}{2} + \frac{3}{4} - \frac{T_2}{2} \tag{6.43}$$

Let $Y_1' = Y_1/2 + 3/4$. Then $Y_1 \in A_1$ is equivalent to

$$Y_1' \in A_1' = \left\{ y \in \mathbb{R} : \frac{3a+10}{8} \leq y < \frac{3a+13}{8} \right\}, \tag{6.44}$$

and (6.43) is transformed into

$$Y_2 = Y_1' - \frac{T_2}{2} \tag{6.45}$$

Let $\psi_y(B)$ be a set mapping such that $\{b \in B\} \rightarrow \{-2b + 2y\}$. (Then $\psi_{Y_1'}(\{Y_1' - T_2/2\}) = \{T_2\}$.) ψ_y is an invertible mapping, so $x \in B$ if and only if $\psi_y(x) \in \psi_y(B)$.

Applying $\psi_{Y_1'}$ to (6.42), we may write,

$$\begin{aligned} & P(Y_2 \in (B \cap Q_a \cap S^0)(0) | Y_1 \in A_1, X_1 = x_1, X_0 = x_0, J_3) \\ =\ & P(Y_1' - \frac{T_2}{2} \in (B \cap Q_a \cap S^0)(0) | Y_1' \in A_1', X_1 = x_1, X_0 = x_0, J_3) \\ =\ & P(T_2 \in \psi_{Y_1'}((B \cap Q_a \cap S^0)(0)) | Y_1' \in A_1', X_1 = x_1, X_0 = x_0, J_3). \end{aligned} \tag{6.46}$$

Here, $(B \cap Q_a \cap S^0)(0) \subset Q_a(0)$, and $Y_1' \in A_1'$ together imply that,

$$\psi_{Y_1'}((B \cap Q_a \cap S^0)(0)) \subset (-0.5, 0.5). \tag{6.47}$$

To see this, we calculate the minimum and maximum values of the mapping $\psi_{Y_1'}(Q_a(0))$, for $Y_1' \in A_1'$:

Recall that

$$Q_a = \left\{ x \in S : \frac{3a+11}{8} \leq x(0) < \frac{3a+12}{8} \right\},$$

so

$$Q_a(0) = \left\{ x \in \mathbb{R} : \frac{3a+11}{8} \leq x < \frac{3a+12}{8} \right\}.$$

Therefore,

$$\begin{aligned} \min(\psi_{Y_1'}(Q_a(0))) &= -2\max(b : b \in Q_a(0)) + 2\min(y : y \in A_1') \\ &= -2\frac{a+12}{8} + 2\frac{a+10}{8} \\ &= -0.5 \end{aligned}$$

$$\begin{aligned} \max(\psi_{Y_1'}(Q_a(0))) &= -2\min(b : b \in Q_a(0)) + 2\max(y : y \in A_1') \\ &= -2\frac{a+11}{8} + 2\frac{a+13}{8} \\ &= 0.5. \end{aligned}$$

Therefore $\psi_{Y_1'}((B \cap Q_a \cap S^0)(0)) \subset \psi_{Y_1'}(Q_a)(0) \subset (-0.5, 0.5)$.

We are ready to invoke Lemma 6.3. We have:

- $J(X_1) = 0$,
- $J(X_2) = 1$,
- $X_1(1) = X_2(0) \in A_1$, as in (6.41),
- A_1 contains no element smaller than $7/4$ (since $a \geq 1$),
- $\psi_{Y_1'}((B \cap Q_a \cap S^0)(0)) \subset (-0.5, 0.5)$.

Therefore, by Lemma 6.3,

$$\begin{aligned} & P(T_2 \in \psi_{Y_1'}((B \cap Q_a \cap S^0)(0))|Y_1' \in A_1', X_1 = x_1, X_0 = x_0, J_3) \\ \geq & \frac{k_{min}}{1 + \frac{1}{2(7/4-1)}} \mu(\psi_{Y_1'}((B \cap Q_a \cap S^0)(0))) \\ = & \frac{3k_{min}}{5} \mu(\psi_{Y_1'}((B \cap Q_a \cap S^0)(0))). \end{aligned} \tag{6.48}$$

The action of ψ_y is to translate a set by $-y$, multiply all its coordinates by -1, and double its size. The first two actions of ψ don't change the measure since the image under translation is guaranteed to be contained in $(-0.5, 0.5)$, by (6.47). The only action that affects the measure is the size doubling. This means,

$$\mu(\psi_{Y_1'}(B \cap Q_a \cap S^0)(0)) = 2\mu((B \cap Q_a \cap S^0)(0)). \tag{6.49}$$

Finally, since μ only measures elements with no jumps, the point in the interval where Lebesgue measure is taken is actually arbitrary, (see Figure 10). Therefore,

$$\mu((B \cap Q_a \cap S^0)(0)) = \mu((B \cap Q_a \cap S^0)(1)) = \mu(B \cap Q_a \cap S^0). \tag{6.50}$$

The last μ is the Lebesgue measure defined on the right-hand endpoint specified in Definition 6.1.

This gives,

$$P(T_2 \in \psi_{Y_1'}((B \cap Q_a \cap S^0)(0))|J_3) \geq \frac{3k_{min}}{5} 2\mu(B \cap Q_a \cap S^0),$$

and hence, by (6.37), (6.38), (6.42), and (6.46),

$$P^3(x_0, B) \geq \alpha\beta \cdot \mu(B \cap Q_a \cap S^0).$$

Therefore, C_a is ν_3^a-small $\forall\ a \geq 1$. ■

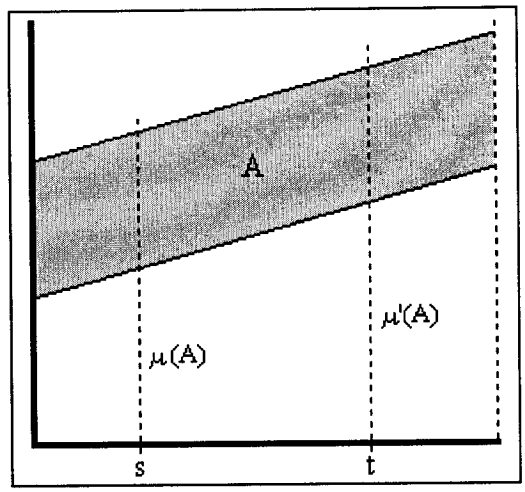

Figure 10: *Lebesgue measure taken at two different points s and t. Since only elements with no jumps are measured, $\mu(A)$ and $\mu'(A)$ must be equal.*

Small sets of the sort above will be important in the next section when we consider aperiodicity. Since small sets are petite, and since a union of petite sets is also petite, we can use their existence to prove Corollary 6.1.

Proof of Corollary 6.1.

Since C is a finite union of petite sets, C_a, $a = 0, 1, 2, ...$, it will automatically be petite if the C_a are. Theorem 6.3 establishes this for $a \geq 1$, so it need only be shown that C_0 is petite.

For the set C_0, the bound of Lemma 6.3 is insufficient. However, C_0 can be shown to be petite by using the petiteness of C_1 and C_2, and noting that any element of C_0 will fall in $C_1 \cup C_2$ after one zero-jump step.

Let $C_{1,2} = C_1 \cup C_2$. Then $C_{1,2}$ is petite with some measure that we will call $\nu^{1,2}$. That is, for $x \in C_{1,2}$ and $B \in \mathcal{B}(S)$, we have,

$$\sum_{n=0}^{\infty} P^n(x, B)a(n) \geq \nu^{1,2}(B),$$

for some measure $a(n)$ over the integers.

Now note that for $x \in C_0$,

$$P(x, C_{1,2}) \geq P(\text{no jumps in } x) = q > 0.$$

The justification for this is similar to that of Lemma 6.1. The number of jumps in x is Poisson distributed with intensity proportional to the area in the previous interval. Since this area is finite, there is a positive probability q that there are no jumps.

Therefore, for any n, any set B, and any $x \in C_0$,

$$P^n(x, B) \geq qP^{n-1}(x+1, B).$$

But $x + 1 \in C_{1,2}$. Let $b(n) = a(n-1)$, with $b(0) = 0$, so that $\{b(n)\}$ is also a distribution on $\mathbb{Z}^+$. Then for all $x \in C_0$,

$$\begin{aligned}
\sum_{n=0}^{\infty} P^n(x, B)b(n) &\geq \sum_{n=0}^{\infty} qP^{n-1}(x+1, B)b(n) \\
&= \sum_{n=1}^{\infty} qP^{n-1}(x+1, B)a(n-1) \\
&= q\sum_{m=0}^{\infty} P^m(x+1, B)a(m) \\
&\geq q\nu^{1,2}(B).
\end{aligned}$$

Therefore, the set C_0 is petite, so C is also petite. ∎

6.3 Aperiodicity

Markov chain aperiodicity is a desirable property because it greatly simplifies the overall behavior of the chain. Without aperiodicity, the chain can never settle down to a steady state, but will move between states in the stochastic version of a limit cycle. Aperiodicity is, of course, essential to the *Aperiodic* Ergodic Theorem, and this section is devoted to its establishment for the Markov chain of window sizes.

Section 5.4.3 of [2] discusses the technical details of periodicity. Some of these results are given below, and are used to prove that the Markov chain of TCP window sizes is aperiodic.

Definition 6.7 *Suppose, for a ψ-irreducible chain, there exists a ν_M-small set C satisfying $\nu_M(C) > 0$. Define*

$$E_C = \{n \geq 1 : C \text{ is } \nu_{\mathrm{n}} - \text{small, with } \nu_{\mathrm{n}} = \delta_{\mathrm{n}}\nu_{\mathrm{M}} \text{ for some constant } \delta_{\mathrm{n}} > 0\}.$$

The set E_C is taken from (5.40) of [2]. The idea is that, since $\nu_M(C) > 0$, there is a positive probability that the chain may return to C in M steps. The set E_C builds on this by guaranteeing that the chain may return to C in n steps, for any $n \in E_C$. The greatest common divisor of E_C, denoted d, is the period of the chain. Therefore, a Markov chain is aperiodic if $d = 1$.

The following theorem (Theorem 5.4.4 of [2]) provides a method of checking aperiodicity.

Theorem 6.4 *Suppose that we have a ψ-irreducible Markov chain on a state space S. Let $C \in \mathcal{B}(S)$ with $\nu_M(C) > 0$ be a ν_M-small set and let d be the greatest common divisor of the set E_C. Then there exist disjoint sets $D_1 ... D_d \in \mathcal{B}(S)$ (a "d-cycle"), such that*

1. *for $x \in D_i$, $P(x, D_{i+1}) = 1$, $i = 0 ... d - 1 (\mathrm{mod}\ d)$;*
2. *the set $N = \left[\cup_{i=1}^{d} D_i\right]^c$ is ψ-null.*

The D_i's thus partition the state space S (according to ψ), and the chain cycles through them in a deterministic manner.

We will now prove through contradiction that the Markov chain under study is aperiodic. The proof uses the above theorem in conjunction with the irreducibility measure φ and the small set C_6 already developed. (See Definition 6.1 and (6.24).)

Theorem 6.5 *The Markov chain of Theorem 5.1 is aperiodic.*

Proof Define, for this particular chain,

$$E_C = \{n \geq 1 : C_6 \text{ is } \nu_{\mathrm{n}} - \text{small, with } \nu_{\mathrm{n}} = \delta_{\mathrm{n}}\nu_3^6 \text{ for some constant } \delta_{\mathrm{n}} > 0.\}$$

Recall that, for any appropriate c, the set

$$C_6 = \left\{x \in S : x(0) < c, \frac{36}{8} \leq x(1) < \frac{42}{8}\right\}$$

is ν_3-small by Theorem 6.3. Note that for $c \geq 5$, we may write

$$C_6 \cap S^0 = \left\{ x \in S^0 : \frac{36}{8} \leq x(1) < \frac{42}{8} \right\},$$

since, in this case, $x(0) < c$ automatically. Recall further that,

$$\nu_3^6(C_6) = \alpha\beta \cdot \mu(C_6 \cap Q_3 \cap S^0),$$

where μ is Lebesgue measure on the right-hand endpoint, and

$$\begin{aligned} Q_3 \cap S^0 &= \left\{ x \in S^0 : \frac{29}{8} \leq x(0) < \frac{30}{8} \right\} \\ &= \left\{ x \in S^0 : \frac{37}{8} \leq x(1) < \frac{38}{8} \right\}. \end{aligned}$$

Clearly, $Q_3 \cap S^0 \subset C_6 \cap S^0$. The intersection $C_6 \cap Q_3 \cap S^0$ is therefore the same as $Q_3 \cap S^0$. Since this set has positive Lebesgue measure, it follows that $\nu_3^6(C_6) > 0$. Therefore E_C may be defined as above with $3 \in E_C$. The greatest common divisor d must then be either 1 or 3.

Assume that $d = 3$ and consider the 3-cycle of Theorem 6.4. The set $(D_0 \cup D_1 \cup D_2)^c$ must be ψ-null. Since φ is absolutely continuous with respect to ψ (see [2], Proposition 4.2.2), the set is also φ-null. We must then have,

$$\varphi(D_0 \cup D_1 \cup D_2) = 1. \tag{6.51}$$

Therefore, by the definition of φ, $D_0 \cup D_1 \cup D_2$ must contain practically all the zero-jump trajectories (less those satisfying $x(0) \leq 1$, see Definitions 6.1 and 5.2).

At the very least, this implies that, without loss of generality on the indices, D_0 contains a set of zero-jump trajectories with positive φ measure. Label this set A_0.

Now, by Theorem 6.4, $P(x, D_1) = 1$ for $x \in A_0$. Since it is certainly possible that there are no jumps in the state following x, we must have that D_1 contains all the zero-jump trajectories that follow from those in A_0. Specifically, D_1 contains the set $A_1 = \{x + 1 : x \in A_0\}$.

Continuing in this fashion, we see that the set of zero-jump trajectories must spiral through the 3-cycle as in Figure 11. Two important aspects of this structure must be highlighted.

- If $x \in A_i$, then $x + 3k \in A_i$ for all integers $k \geq 1$.
- Theorem 6.4 states that the sets D_i are disjoint. In particular the sets A_i must also be disjoint.

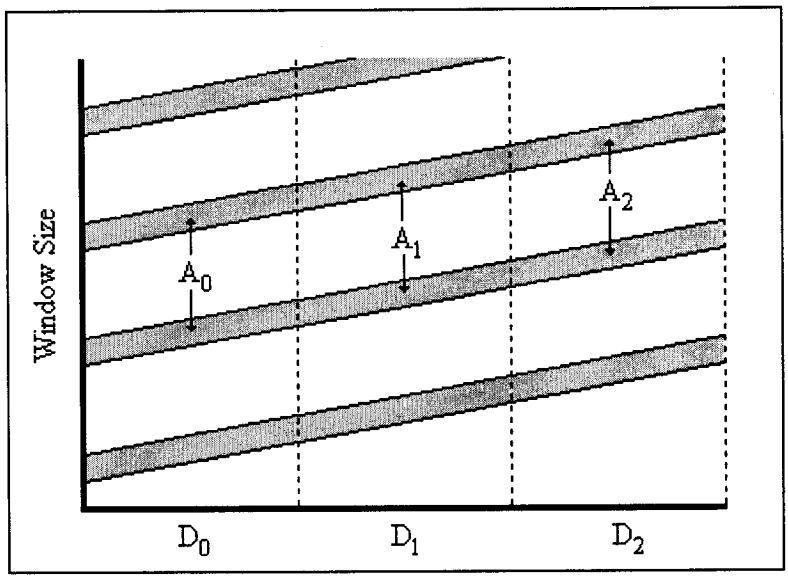

Figure 11: *If d=3, the set of all zero-jump trajectories is divided between D_0, D_1, and D_2. Of course, these sets need not be intervals.*

There is nothing contradictory about the structure above. The contradiction follows when jumps are introduced into the Markov chain, and the disjointness condition is violated.

To proceed with the contradiction, re-label (without loss of generality) D_0, D_1, D_2 so that $A_0 \cap \{x \in S^0 : 2 < x(0) < 3\}$ has positive measure (with respect to Lebesgue measure on the right-hand endpoint). This is possible since the D_i's cover all but a null set of trajectories. Let

$$A_0^{sub} = A_0 \cap \{x \in S^0 : 2 < x(0) < 3\}. \tag{6.52}$$

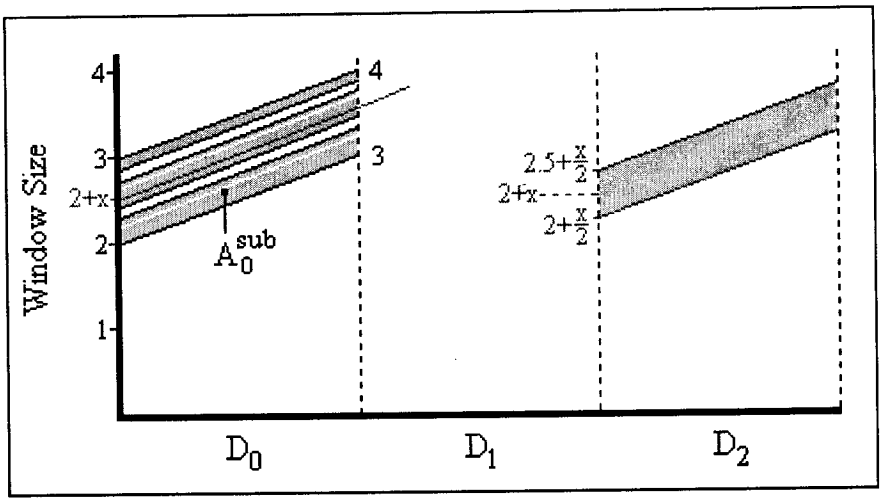

Figure 12: *In a 3-cycle, D_0 and D_2 overlap, leading to a contradiction.*

Starting from any element in A_0^{sub}, there is a positive probability that a single jump will occur in the next step of the chain. By Theorem 6.4, D_1 must therefore contain all trajectories with one jump and left-hand endpoint in $A_0^{sub}(1)$.

Consider any element $a \in A_0^{sub}$. a has left-hand endpoint $2 + x$ and right-hand endpoint $3 + x$ for some $0 < x < 1$. By (6.9), D_1 then contains the element with left-hand endpoint $3 + x$ and right-hand endpoint lying in the continuous interval $(2 + x/2, 2.5 + x/2)$. (See Figure 12.)

Of course, starting from here, there is a positive probability that there will be no jumps in the next step, so invoking Theorem 6.4 again leads to the conclusion that D_2 must contain the set of trajectories with no jumps and left-hand endpoints within $\left(2 + \frac{x}{2}, 2.5 + \frac{x}{2}\right)$.

A simple calculation shows that, for $0 < x < 1$,

$$2 + \frac{x}{2} < 2 + x < 2.5 + \frac{x}{2}. \tag{6.53}$$

Since both contain the arbitrary element a, A_0^{sub} and D_2 are not disjoint. Hence D_0 and D_2 are not disjoint. This violates Theorem 6.4, leading to a contradiction of the assumption that $d = 3$.

Therefore, we must have that $d = 1$, making the chain aperiodic. ∎

There is no natural periodicity in the evolution of window sizes, so the preceding result is not surprising. However, establishing aperiodicity is an essential step in establishing the existence of a unique stationary distribution, so the exercise was necessary.

All the properties established above (ψ-irreducibility, the existence of petite sets, and aperiodicity) indicate that the general state-space Markov chain is, in a sense, well behaved. Ergodic theorems for Markov chains generally rely on such properties, and the Aperiodic Ergodic Theorem is no exception. The next section investigates whether this "well-behaved" chain is actually stable by observing the mean behavior of the chain over time.

6.4 The Foster-Lyapunov Drift Criteria

Armed with a φ-irreducible, aperiodic chain and a petite set, the final step to satisfying the hypotheses of Theorem 6.1 is the discovery of a function satisfying (6.1), the equation known as the Foster-Lyapunov drift criterion. This criterion, if satisfied, implies that the chain returns to a petite set, the general argument being that (6.1) can be used to show that the mean return time to C is finite. In order to show that the drift criterion is satisfied, the following definitions, found in [2], must be used.

Definition 6.8 *For some test function $V(x) : \mathcal{B}(S) \rightarrow \mathbb{R}$, Define $\Delta V(x)$ to be the one-step mean drift of a Markov chain by:*

$$\Delta V(x) = \int_S P(x, dy)V(y) - V(x) \tag{6.54}$$

This is interpreted as the expected value of the change in the function V after one transition of the Markov chain.

Definition 6.9 *The Foster-Lyapunov drift criterion is said to be satisfied if there is a non-negative, finite test function V, a finite b, and a petite set C such that:*

$$\Delta V(x) \leq -1 + b\chi_C(x), \ x \in S \tag{6.55}$$

The Foster-Lyapunov criterion is reminiscent of the Lyapunov stability condition for dynamical systems. In both cases, there is a positive function V which decreases when acted upon by the system dynamics. This V can be thought of as representing the "energy" of the system. Since the energy tends to decrease as the system evolves, the system must eventually reach some sort of minimal-energy state. One notable difference from the classical Lyapunov condition is that the system can escape from this minimal state since $\Delta V(x)$ is allowed to be positive on C. The point, however, is that once outside C, the mean return time to C is finite.

If such a function can be found for the Markov chain of window sizes, then stochastic stability will be close at hand. The following theorem uses a simple function, which satisfies the drift criteria not on the petite set C, but on a related "gateway" set called G. The fact that trajectories in G are guaranteed to reach C in the next step may then be used along with the drift condition as an equivalent to the original Foster-Lyapunov criteria.

Theorem 6.6 *Assume that there is a minimum loss probability $k_{min} > 0$ at the router. Let*

$$h(z) = \frac{\ln(\frac{z}{z-4})}{z - 0.5}, \; z > 4.$$

If $c \geq h^{-1}(k_{min})$, and $c \in \{5, 8, 11, ...\}$, then the function $V(x) = x(1)$ satisfies the Foster-Lyapunov drift criteria for the Markov chain defined by Theorem 5.1 for the (not necessarily petite) set

$$G = \{x \in S : x(1) < c\}.$$

Furthermore, If C is the petite set $\{x \in S : x(0) < c\}$ for the same c as G, then for $x \in G$, $P(x, C) \equiv 1$, and for $x \in G^c$, $P(x, C) \equiv 0$.

It will be shown that satisfying the above theorem is equivalent to satisfying the drift criteria directly on a petite set. Essentially it asserts that, outside a petite set C, the chain drifts toward a gateway set, from which C is immediately reached in one step. Since the sole purpose of the drift criteria is to establish that the mean time to return to C is finite, this modification should be intuitively satisfactory.

Proof First note that $V(x)$ is non-negative everywhere, since the chain takes on only non-negative values.

Let $x, y \in S$ and define $t = 0$ at the start of state x. Recall that the window size can grow by at most one unit per step of the chain. If the chain takes x to y in one step, then $V(y) = y(1) \leq x(1) + 1$, where $x(1)$ and $y(1)$ are the right-hand endpoints of states x and y. Therefore, $\forall\ x$,

$$\int_S P(x, dy)V(y) \leq x(1) + 1. \tag{6.56}$$

This leads to,

$$\begin{aligned} \Delta V(x) &= \int_S P(x, dy)V(y) - V(x) \\ &\leq x(1) + 1 - x(1) = 1. \end{aligned} \tag{6.57}$$

This holds for both $x \in G$ and $x \in G^c$, so we may fix the value of b in (6.55) to be 2 as long as it can be shown that, for $x \in G^c, \Delta V(x) \leq -1$.

For $x \in G^c$ (and indeed for all $x \in S$) the stochastic intensity of the jump process in the next step y is given by (5.11) to be proportional to the area under the trajectory x. For a given value of $x(1)$, this area is bounded below by the case where $J(x) = 0$. In this case, $x(t) = x(1) - 1 + t,\ 0 \leq t < 1$, and hence,

$$\begin{aligned} \Lambda(1, 2) &= \int_0^1 k(\tau_n(u))x(u)du \\ &\geq \int_0^1 k_{min}(x(1) - 1 + u)du \\ &= k_{min}(x(1) - 0.5) \end{aligned}$$

Therefore, since the jump probability is Poisson,

$$p_0 = P(\text{no jumps in state y}) = \mathrm{e}^{-\Lambda(1,2)} \leq \mathrm{e}^{-\mathrm{k_{min}(x(1)-0.5)}} \tag{6.58}$$

Now, as a first approximation, the state y can either either contain no jumps or at least one jump. If there are no jumps, then $y(1) = x(1) + 1$. Otherwise, $y(1)$ must be $x(1)/2+1$ at a maximum, since this is the highest value attainable under a single jump (see (6.9)), and more jumps will drive $y(1)$ even lower. (See Figure 13.) Therefore,

$$V(y) \leq (x(1) + 1)\chi_{\{J(y)=0\}} + \left(\frac{x(1)}{2} + 1\right)\chi_{\{J(y)\neq 0\}} \tag{6.59}$$

In (6.54), this approximation yields:

$$\begin{aligned}
\Delta V(x) &\leq \int_S P(x,dy)\left((x(1)+1)\chi_{\{J(y)=0\}} + \left(\frac{x(1)}{2}+1\right)\chi_{\{J(y)\neq 0\}}\right) - x(1) \\
&= \int_{S^0} P(x,dy)(x(1)+1) + \int_{(S^0)^c} P(x,dy)\left(\frac{x(1)}{2}+1\right) - x(1) \\
&= p_0(x(1)+1) + (1-p_0)(\frac{x(1)}{2}+1) - x(1) \\
&= 1 - \frac{x(1)}{2}(1-p_0) \qquad (6.60)
\end{aligned}$$

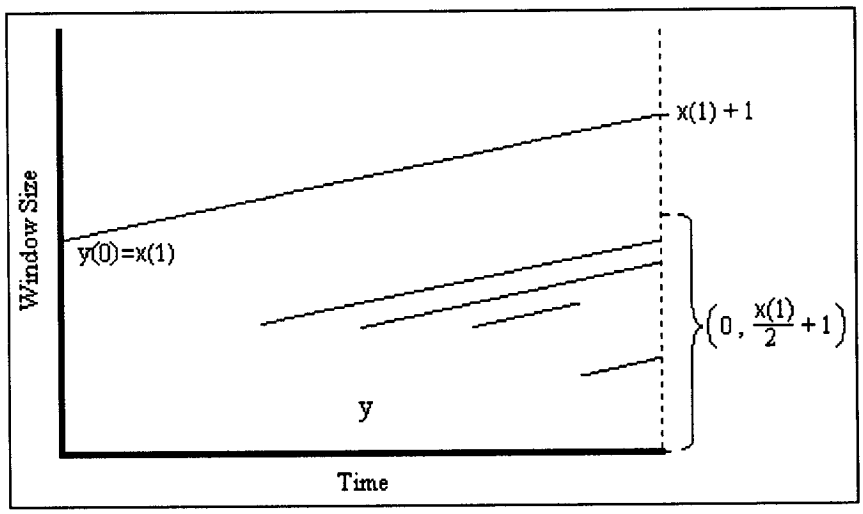

Figure 13: *The possible values of $y(1)$ based on whether or not at least one jump occurs.*

For small values of $k(t)$, losses occur infrequently. This means that there will rarely be more than one jump in a round-trip time, and the approximation above will be good.

For $x \in G^c$, the criterion $\Delta V(x) \leq -1$ can now be satisfied if

$$\begin{aligned} -1 &\geq 1 - \frac{x(1)}{2}(1-p_0) \\ \text{or if, } 4 &\leq x(1)(1-p_0) \\ \text{or if, } 4 &\leq x(1)(1-e^{-k_{min}(x(1)-0.5)}) \\ \text{or if, } \mathrm{k_{min}} &\geq \frac{\ln(\frac{x(1)}{x(1)-4})}{x(1)-0.5} \end{aligned} \tag{6.61}$$

with the restriction $x(1) > 4$.

The right side of (6.61) is the function $h(x(1))$. This can be shown to be a decreasing function of $x(1)$ when $x(1) > 4$, with range $(0,1)$. (Hence h is indeed invertible on its restricted domain.) But for $x \in G^c$, it is automatically true that $x(1) \geq c \geq h^{-1}(k_{min})$. Therefore,

$$\begin{aligned} h(x(1)) &< h(c) \\ &\leq h(h^{-1}(k_{min})) \\ &= k_{min}. \end{aligned}$$

That is, $\forall\ x \in G^c$, (6.61) is satisfied, so $\Delta V(x) \leq -1$.

Finally, for $x \in G$, $x(1) < c$, and therefore $y(0) < c$, which implies that $y \in C$. Thus for such trajectories, the chain enters into C in a single step with probability one. Conversely, if $y \in C$, then $y(0) < c$ and hence $x(1) < c$ so that $x \in G$. In other words, only trajectories in G can lead to C, so there is probability zero that the chain enters C from G^c. ■

One can always find sets G and C satisfying $c \geq h^{-1}(k_{min})$, by simply taking c large enough. Indeed, as k_{min} becomes smaller, the sets becomes larger. This is to be expected since a smaller loss probability $k(t)$ leads to a steady-state with a wider distribution of window sizes (see Theorem 2.1 and (2.12)). In any case, the chain tends toward G. Intuitively, copies of the process which drift into this set will, in the next step, drift into the petite set C, where they may be allowed to couple, resulting in a steady state.

To show rigourously that drift toward G is sufficient, we must re-visit and expand upon Theorem 13.0.1 of [2].

Definition 6.10 *Define τ_A to be the first return time to A, given by:*

$$\tau_A = min\{n \geq 1 : X_n \epsilon A\}.$$

This definition is also due to [2].

Theorem 13.0.1 states (in part) the following:

Theorem 6.7 *(Aperiodic Ergodic Theorem II)*

For an aperiodic Harris recurrent chain with an invariant measure π, if there exists a petite set C such that

$$\sup_{x \epsilon C} E_x(\tau_C) < \infty,$$

then there is a unique invariant probability measure π such that, for every $x \epsilon S$,

$$\sup_{A \epsilon \mathcal{B}(S)} |P^n(x, A) - \pi(A)| \to 0 \text{ as } n \to \infty.$$

This gives an equivalent criterion to the drift criterion of Theorem 6.1, and hence implies that the Foster-Lyapunov drift criteria is satisfied for some unknown function $V(x)$ and the set C.

To show that the result of Theorem 6.7 holds, it must now be shown that the chain is both Harris recurrent and that, starting in C, the mean return time to C is finite (the existence of an invariant measure follows automatically from Harris recurrence via Theorem 10.4.4 of [2]). The following lemma, which builds on Theorem 6.6 is necessary for satisfying the hypotheses of Theorem 6.7.

Lemma 6.4 *For all $x \epsilon S$, the petite set C with $c \geq h^{-1}(k_{min})$, and the corresponding gateway set G,*

$$E_x[\tau_C] \leq x(1) + 1 + 2\chi_G(x).$$

Proof For any $x \epsilon S$ let $X_0 = x$. Starting from any x, define τ_C to be the first $k \geq 1$ such that $X_k \epsilon C$. Since C is reached only through G, it must be true that $X_{\tau_C - 1} \epsilon G$. What is more, by definition of the return time, it is also true that $X_k \epsilon G^c$ for $1 \leq k < \tau_C - 1$, since otherwise C would be reached before $k = \tau_C$.

Take the test function $V(X_{k-1}) = X_{k-1}(1)$. By Theorem 6.6,

$$1 \leq -\Delta V(X_{k-1}) + 2\chi_G(X_{k-1}). \tag{6.62}$$

Summing (6.62) from one to $\tau_C - 1$ then yields,

$$\sum_{k=1}^{\tau_C-1} 1 \leq \sum_{k=1}^{\tau_C-1} (-\Delta V(X_{k-1}) + 2\chi_G(X_{k-1})). \tag{6.63}$$

But for $1 \leq k < \tau_C - 1$, $\chi_G(X_k) = 0$. Therefore,

$$\sum_{k=1}^{\tau_C-1} 1 \leq -\sum_{k=1}^{\tau_C-1} \Delta V(X_{k-1}) + 2\chi_G(X_0). \tag{6.64}$$

Now, by definition,

$$\Delta V(X_{k-1}) = E[V(X_k|X_{k-1})] - V(X_{k-1}). \tag{6.65}$$

Substituting (6.65) into (6.64) and taking expectations on both sides yields,

$$\begin{aligned} E_x[\tau_C] - 1 &\leq E_x\left[-\sum_{k=1}^{\tau_C-1} E[V(X_k|X_{k-1})] - V(X_{k-1})\right] + 2\chi_G(x) \\ &= -\sum_{k=1}^{\tau_C-1} E_x[V(X_k)] - E_x[V(X_{k-1})] + 2\chi_G(x) \\ &= -E_x[V(X_{\tau_C-1})] + E_x[V(X_0)] + 2\chi_G(x) \\ &\leq V(x) + 2\chi_G(x). \end{aligned} \tag{6.66}$$

Since $V(x) = x(1)$, the result follows directly. ∎

The hypotheses of Theorem 6.7 can now be satisfied. For any $x \in C$, we have $x(1) < \infty$ so that,

$$E_x[\tau_C] \leq x(1) + 1 + 2\chi_G(x) < \infty \tag{6.67}$$

by Lemma 6.4. Thus $\sup_{x \in C} E_x[\tau_C] < \infty$. Moreover, since $E_x[\tau_C] < \infty$ for all $x \in S$, it follows that $P_x(\tau_C < \infty) = 1$ for all $x \in S$. By [2], Theorem 9.1.3(ii), the chain is therefore Harris recurrent.

The arguments above provide a small innovation to the Aperiodic Ergodic Theorem. The chain need not drift explicitly toward a petite set if there is a series of suitable "gateway" sets that act as a funnel for the chain. Of course, the end result implies that there is a test function and a petite set satisfying the drift criteria directly, and hence a more inspired look at the problem might reveal such entities. On the other hand, Lyapunov functions are notoriously difficult to find, and any mechanism that eases the burden of searching for them is welcomed.

6.5 Summary

The Markov Chain describing the evolution of a TCP connection's window size has now been shown to satisfy all the hypotheses of the Aperiodic Ergodic Theorem. The Markov chain is irreducible and aperiodic, and tends toward a well-defined pseudo-atomic set. The consequence is that there is a unique invariant measure π such that, regardless of the starting point, the long-run probability of the chain being in any set A tends to $\pi(A)$, (see (6.2)).

Elements of such sets correspond not to a particular window size, but to a function describing the window size of a connection over an entire round-trip time. The measure π on the state-space S of such elements is incredibly complicated, and no attempts will be made to show it explicitly. However, the fact remains that, for an individual TCP connection on its own time scale, the value of the window size tends to a unique steady state which is determined solely by the packet loss probability.

Chapter 7

Conclusions

The stochastic stability of a single TCP source, while interesting, has only limited relevance on its own. This section concludes the discussion by considering the consequences of stochastic stability for an entire network of TCP sources. It will be shown that the stability of individual TCP sources suggests the stability of an entire network when the loss probability is chosen appropriately. Some criteria for choosing such a loss probability are given, and simulation results are shown that provide specific examples of network stability.

7.1 Limiting Behavior for Many Sources

In the last chapter we established that the window size of a single TCP connection converges to a unique steady-state behavior on a local time scale when subjected to a constant loss probability. Consider now a large network of N TCP sources observed after a long time such that each window size trajectory has reached steady state and is distributed according to the distribution π given in the last chapter. It is natural to suspect that since each window size trajectory is in steady state, the collective window size distribution of the entire network is also in steady state. By uniqueness, such a steady state could correspond to the one characterized in Theorem 2.1, and would thus be fully known. Furthermore, a steady-state distribution of window sizes should also coincide with a steady-state queue size, since the rate of packet arrivals

to the router queue depends on the window sizes. In this manner, a constant drop probability should lead not just to stable window size trajectories, but to the stability of an entire network.

While this reasoning is certainly plausible and consistent with numerical results, it is difficult to show conclusively. One major difficulty arises from the fact that π is a measure on trajectories, while a better notion of stability would require a fixed-point distribution for *instantaneous* window sizes. A second difficulty is that it is not immediately clear that the queue size is in steady state merely because the window sizes are. This section explores these issues and makes arguments in favor of network stability, where the network is represented by the mean-field model of [1].

Let $X_{i,n}$ be the i^{th} state of the Markov chain of window sizes for the n^{th} TCP connection out of a possible N connections. Let $\{V_n,\ n = 1 \ldots N\}$ be random variables on S with distribution π. Since the random variables $X_{i,n}$ are eventually governed by π, the V_n can be thought of as representing the $X_{i,n}$ in equilibrium. The independence of these V_n is formalized in the following lemma.

Lemma 7.1 *If the random variables V_n are taken to represent canonical window sizes in equilibrium, then $\{V_n,\ 1 \leq n \leq N\}$ are independent and identically distributed with distribution π.*

Proof By definition, each V_n is distributed according to π, which is unique according to the last chapter. Independence can be shown by reasoning that the window sizes interact only through the loss rate and the round-trip time (see (2.1) and (2.2)). Since the loss rate is constant, no interaction can arise in this way. Also, since each V_n represents a canonical window size, the round-trip times are normalized to one, and no interaction can occur in this way either. ■

As $i \to \infty$, the collective distribution of the $X_{i,n}$ thus converges to a steady state which is the product measure of N copies of π. In other words the N canonical window size trajectories are asymptotically independent with distribution π.

If the V_n represented instantaneous window sizes, the result would be immediate: by the Law of Large Numbers, a large network of TCP sources will eventually be

distributed according to π because each individual window size will become i.i.d. with distribution π. (In taking the limit, the link rate per connection, C, is held constant, so the true link rate L scales up with N since $L = NC$.) More formally, for any Borel set $A \in \mathcal{B}(S)$, define the indicator function $\chi_{V_n}(A)$ so that $E(\chi_{V_n}(A)) = \pi(A)$. For any A, each $\chi_{V_n}(A)$ is an i.i.d. random variable with finite mean, so the Strong Law of Large Numbers gives,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \chi_{V_n}(A) = \pi(A), \ a.s. \tag{7.1}$$

In other words, the proportion of sources in steady state with window sizes in a set A would be $\pi(A)$. Furthermore, since the window sizes are instantaneous, π would coincide with the measure in Theorem 2.1 by uniqueness. Unfortunately, the V_n represent window size trajectories over a full round-trip time, so this is not the case. On the other hand, one can argue that round-trip times are short, so each V_n is almost instantaneous. From this point of view, the steady-state result above should hold at least approximately.

To understand the problems that arise due to V_n being a trajectory, imagine that a large number of window sizes are sampled at time t. This time is sufficiently large that all the window size trajectories are in steady state. Since each V_n resides on its own time scale, an instantaneous sample returns values $V_n(\rho_n^N(t))$ for each n, where $\rho_n^N(t)$ denotes the rtt time of source n out of N sources at t seconds (ρ_n^N is the same as the time change τ_n^{-1} defined in Chapter 4, but for a specific number of sources). These $\rho_n^N(t)$ are random variables for finite N. They have different values, but are in fact strongly dependent on each other because the round-trip times that determine them are related through a common queueing delay. The window size values returned by sampling at any time t are thus not independent, so the Law of Large Numbers no longer guarantees a network steady state, even though the canonical window size trajectories are i.i.d.

One observation that simplifies matters and may help resolve this problem is that the time-instantaneous marginals of π, given by

$$\pi_t(B) = \pi(x \in S : x(t) \in B), \ B \in \mathcal{B}(\mathbb{R}), \tag{7.2}$$

appear to be identical for all values of t. This claim is supported by the fact that the canonical process X_i could have, for example, been started at time 1/2 instead of time 0 but the limit would be the same. In other words the stationary distribution is independent of a time shift, and hence the marginals are identical (this property can be attributed to the randomness of the jump times).

Since the V_n are identically distributed, and since the observation above shows that the distribution of $V_n(\rho_n^N(t))$ is independent of the actual $\rho_n^N(t)$, the instantaneous window sizes might also be independent in some form. Such independence would allow the Law of Large Numbers to again be invoked, giving a steady-state result for instantaneous window sizes. (One must be careful to note that it is not true that $V_n(\rho_n^N(t))$ has a distribution equal to the marginal of π. There is a correlation between the window size in rtt time and the $\rho_n^N(t)$. It is only in the limit as $N \to \infty$ that $\rho_n^N(t) \to \rho_n(t)$ where $\rho_n(t)$ is a deterministic function of the time t in seconds that $V_n(\rho_n(t))$ has the distribution π_t.)

The real-time instantaneous window sizes are not independent in general, but they do become so once they reach a steady state. This fact is proven in [11] for the mean-field limit. It may also be seen by considering Theorem 2.1, which shows that, in steady state, the window size distribution is independent of the round-trip time, so the sole source of interaction (for constant k) disappears. This suggests that the interaction between instantaneous window sizes must be weak when each individual connection, along with the queue size, is close to steady state. This is notable because mean-field models are generally built specifically because interaction exists, and the fact that it disappears in steady state shows that such a steady state may occur without it.

The arguments above are made to suggest that the instantaneous window sizes of TCP connections in a bottleneck network can eventually be described by a collection of i.i.d. random variables. While the details are not fully worked out here, this argument leads to the following conjecture by virtue of the Law of Large Numbers:

> A large number of sources subject to a constant packet drop probability are collectively governed by a steady state window size distribution after a sufficient amount of time.

We take this to be given for the rest of this section.

The second problem is that, even if the distribution of window sizes converges a steady state, it is unclear that the queue size and round-trip times should follow suit. Although it is likely that this is the case, such a result is beyond the scope of this thesis and is left as a conjecture. Below we merely present an existence theorem for such a steady state when the packet loss rate is constant, provide some theoretical plausibility for its stability, and state that the numerical evidence also supports stability.

Theorem 7.1 *When the packet loss probability is constant, there exists a steady state for the mean-field limit described by (2.6) and (2.7).*

Proof Let $W_n^N(t)$, $n = 1, 2, \ldots, N$ and $Q^N(t)$ denote the solution to the finite system with N sources. One can prove the existence (see [11]) of an almost-sure mean-field limit $W_n(t)$, $n = 1, 2, \ldots, \infty$ and $Q(t)$ when $N \to \infty$. Let T_n be the transmission delay of connection n. Suppose that sources can be divided into d classes each with the same transmission delay so class c has transmission delay T_c. The round-trip time of class c is then $R_c(t) = T_c + Q(t - R_c(t))/L$, as in (2.4). Let the proportion of sources in class c among the first N sources be κ_c^N. We suppose $\lim_{N\to\infty} \kappa_c^N \to \kappa_c$.

Consider the equation,

$$(1-k)\sum_{c=1}^{d} \kappa_c \frac{m}{T_c + Q/L} = L$$

where m is the expected window size in steady state when the loss probability is k (m is the mean window size according to π). The left hand side is monotonically decreasing in Q so there exists a unique solution Q_{equ}.

Let $\rho_n^N(t)$ denote the rtt time of source n at t seconds and let $X_n(s)$ denote the window size of source n in rtt time (these are independent and don't depend on N). Hence, $W_n^N(t) = X_n(\rho_n^N(t))$. Denote the mean field limit of $\rho_n^N(t)$ by $\rho_n(t)$ so $W_n(t) = X_n(\rho_n(t))$. Note that $\rho_n(t)$ is deterministic since it is a mean field limit.

Now start each source $X_n(s)$ off in steady state over an interval $[-1, 0]$ in rtt time. The existence of this steady state was proven in Chapter 6. Also suppose the queue

is constant and $Q(t) = Q_{equ}$ for $t \leq 0$, so $\rho_n(t)$ is linearly increasing on $[\rho_n^{-1}(-1), 0]$ because the queue was constant so the round-trip times are constant. Hence, over this interval, $W_n(t) = X_n(\rho_n(t))$ is in steady state.

Now, at any time t, if the queue size is measured in non-relative terms and is nonzero, (2.4) and (2.5) give

$$\frac{dQ^N(t)}{dt} = (1-k)\frac{1}{N}\sum_{n=1}^{N}\frac{W_n^N(t)}{T_n + Q(t - R_c(t))/L} - L.$$

This equation holds in the mean field limit and at time $t = 0$ we get

$$\begin{aligned}\frac{dQ(t)}{dt} &= (1-k)\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{W_n^N(t)}{T_n + Q(t - R_n(t))/L} - L \\ &= (1-k)\sum_{c=1}^{d}\kappa_c\frac{m}{T_c + Q_{equ}/L} - L \\ &= 0.\end{aligned} \tag{7.3}$$

Consequently the mean field queue stays constant at Q_{equ}. Since the queue remains constant, the round-trip times don't change so $W_n(t) = X_n(\rho_n(t))$ remains in steady state. We conclude that $W_n(t)$, $n = 1, 2, \ldots, \infty$ and $Q(t)$ has a steady state. ■

From the results in [1], the (unique) marginal distribution of $W_n(t)$ in steady state must have a particular density given by Theorem 2.1. The steady-state queue size is also given by solving (2.4) and (2.13).

The result shows the existence of a steady state, but does not show that it is stable. Suppose we start $X_n(t)$, $n = 1, \ldots, \infty$ off in steady state but $Q(0) \neq Q_{equ}$. Although the canonical window size trajectories converge to a steady state, and the overall window size distribution is suspected of doing so, convergence of the queue size is not guaranteed by the theorem above.

However, one can at least see that convergence is plausible by reasoning as follows. Consider (7.3) again, but with just one class of round-trip times. That is,

$$\frac{dQ(t)}{dt} = (1-k)\frac{m}{T + Q(t - R_n(t))/L} - L.$$

If $Q(t-R_n(t)) < (1-k)m-TL$, then the queue size will be increasing one round-trip time later, while if $Q(t-R_n(t)) > (1-k)m-TL$, the queue size will be decreasing. Hence the queue size naturally recovers from any deviations, and so tends to at least stay close to its steady state value.

In fact, one can say even more than this in the following example. Suppose that, at time s, the queue size has remained constant at some value βQ_{equ} for the past round-trip time, where $Q_{equ} = (1-k)m - TL$. Then for all times $t \leq s + R_n(s)$, we have

$$R_n(t) = T + \frac{\beta Q_{equ}}{L}.$$

We can now calculate the queue size one round-trip time after s as

$$\begin{aligned}
Q(s+R_n(s)) &= Q(s) + \int_s^{s+R_n(s)} \frac{dQ(t)}{dt} dt \\
&= \beta Q_{equ} + \int_s^{s+R_n(s)} \left(\frac{(1-k)m}{T + \beta Q_{equ}/L} - L \right) dt \\
&\approx \beta Q_{equ} + R_n(s) \left(\frac{(1-k)m}{T + \beta Q_{equ}/L} - L \right) \\
&= \beta Q_{equ} + (1-k)m - \left(T + \frac{\beta Q_{equ}}{L} \right) L \\
&= \beta Q_{equ} + (1-k)m - TL - \beta Q_{equ} \\
&= Q_{equ}.
\end{aligned}$$

In this case, the queue has reached its equilibrium value in one round-trip time! Although it will not stay there, this example shows a strong tendency toward equilibrium when the window size distribution is constant, and strengthens the claim above.

Numerical results support this reasoning as well; in simulation the window size distribution and queue size always appear to reach steady state together. Numerous attempts in Matlab to discover a scenario in which the window size distribution is constant while the queue size varies have failed. In other words, stability of window sizes strongly suggests stability of the queue size. This leads us to conjecture that the queue size converges to a steady state when the loss rate is held constant, even when it does not start out in steady state.

Indeed, extensive numerical simulation performed in Matlab shows long-term stability of both the window size distribution and of the queue size when the loss probability is held constant. This evidence, coupled with the mathematical ideas above, lends a strong plausibility to the following result.

> In a TCP network with a large, constant number of connections and any constant round-trip time and bandwidth, there exists a constant loss probability k such that a router which randomly drops each incoming packet with probability k will allow the network to reach a stable steady state. The steady-state window size distribution and queue size are known, and given by Theorem 2.1 and (2.12).

This is the final theoretical result. Only a few practical details remain.

7.2 Choosing the Packet Dropping Probability

The practical problem of finding the appropriate loss probability k deserves further discussion. In order to effectively implement a random drop policy, we must impose three requirements on the number of packets in the network. First note that, for N sources, the total number of packets in the network in steady state is equal to the number of sources times the mean window size. The requirements are:

1. The number of packets in the network must be large enough so that the mean window size is significantly greater than five. This is necessary to ensure that the TCP sources can operate effectively in congestion avoidance, and that timeouts are minimized. (A window size less than four causes a source to time out when a packet is lost, rather than halving its window size and re-transmitting the lost packet.)

2. The number of packets in the network should exceed the external delay-bandwidth product. Let T represent the external delay (the round-trip time minus the queueing delay). Then the external delay-bandwidth product, TL, represents the number of packets that can fit on the "wires" of the network. If the number

of packets is less than this, then full network utilization is not achieved, and resources are being wasted. If the number of packets is larger than TL, then there will be some packets enqueued at the router.

3. The number of packets in the network should not exceed the maximum total delay-bandwidth product. That is, there should not be so many packets that the queue overflows. If this were the case, incoming packets would be dropped with probability one, and control over the loss rate would be compromised. Mathematically, the number of packets should be less than $(T + Q_{max}/L)L$, where Q_{max}/L is the queueing delay due to a full queue.

Recall that, in steady state, the loss probability is directly related to the mean window size via (2.12):

$$\int_w w f_k(w) dw = \alpha\sqrt{\frac{1}{k}}$$

where $\alpha \approx 1.310$. Therefore, the number of packets in the network is given by $N\alpha\sqrt{1/k}$.

k must therefore satisfy the following set of equations.

$$\begin{aligned}
\alpha\sqrt{\frac{1}{k}} &> 5 \\
\alpha\sqrt{\frac{1}{k}} &> \frac{TL}{N} \\
\alpha\sqrt{\frac{1}{k}} &< \frac{(T + \frac{Q_{max}}{L})L}{N} = \frac{TL}{N} + \frac{Q_{max}}{N}.
\end{aligned}$$

This can be reduced to the single condition,

$$\frac{\alpha^2 N^2}{(TL + Q_{max})^2} < k < \min\left(\frac{\alpha^2}{25}, \frac{\alpha^2 N^2}{(TL)^2}\right). \tag{7.4}$$

(7.4) cannot be satisfied for every combination of T, L, Q_{max} and N. In particular, if N becomes too large, then the factor $N/(TL + Q_{max})$ will increase above $1/5$, leaving no solution to (7.4). In this case, there are so many sources that even a very small mean window size will result in a large number of packets in the network, and

the balance between sufficient window sizes and a non-overflowing queue will not be achievable. This places a practical limit on the number of sources N which was not seen in previous sections because L was scaled up with N as $L = NC$. However, this limitation is not of paramount concern because the Law of Large Number convergence of Section 7.1 has been observed through simulation to be very fast. For example, a T3 data link with an external delay of 60 ms. and just 50 sources will have a window size distribution that closely resembles the limiting distribution.

Another observation is that, if the link rate was scaled up with N but Q_{max} was held finite, the range of acceptable k values would decrease as N increased. In the limit, the value $k = \alpha^2/(TC)^2$ would be the only solution to (7.4).

If one knows the maximum and minimum number of network connections, and the worst-case delays, studying (7.4) can lead to some valuable insight into the appropriate choices for k and Q_{max}. It may even be adapted to provide guidelines for choosing the parameters of RED. In fact, the RWFD algorithm is designed to follow these guidelines automatically. If $Q(t) \approx Q_{target}$ then the equilibrium loss probability given by (3.8) automatically satisfies (7.4). This is evident in the results of the next section. Moreover, one can show that there is always a range of choices for Q_{target} which guarantee the RWFD loss probability satisfies (7.4) regardless of the actual queue size.

7.3 Simulation Results for RWFD

In this section we present simulation results for the RED Without Feedback Delay algorithm of Chapter 3. The simulations were carried out using OPNET. All simulations consist of a number of TCP connections sharing the same T3 (45 Mbps) data link. Packets lengths are fixed at 536 bytes, making the link rate L equal to 10433 packets per second. The TCP sources are generally modelled as greedy FTP connections, so that there are always packets waiting to be sent. There are also a small number of HTTP connections, which send data bursts of between one and five seconds long at random times. (These HTTP connections may be regarded as noise, and in the quantity present do not significantly change the results.) In order to study

a scenario in which stability is difficult to achieve, the transmission delays are taken to be quite large. Smaller delays can achieve similar results, but queue size may need to be larger.

The router implementing the random drop algorithm is located at the input to the bottleneck link. The objective is to achieve a constant loss probability and queue size at this router, so that the network becomes well behaved.

The first example involves 300 FTP connections in a network with an external delay of $T = 300$ ms. The router has a maximum buffer size of 5000 packets, and the RWFD algorithm attempts to stabilize the queue to a constant target queue size of 1500 packets. In order to accomplish this, the equilibrium loss probability k_{equ} is calculated every second. The simulation results are shown in the figures below.

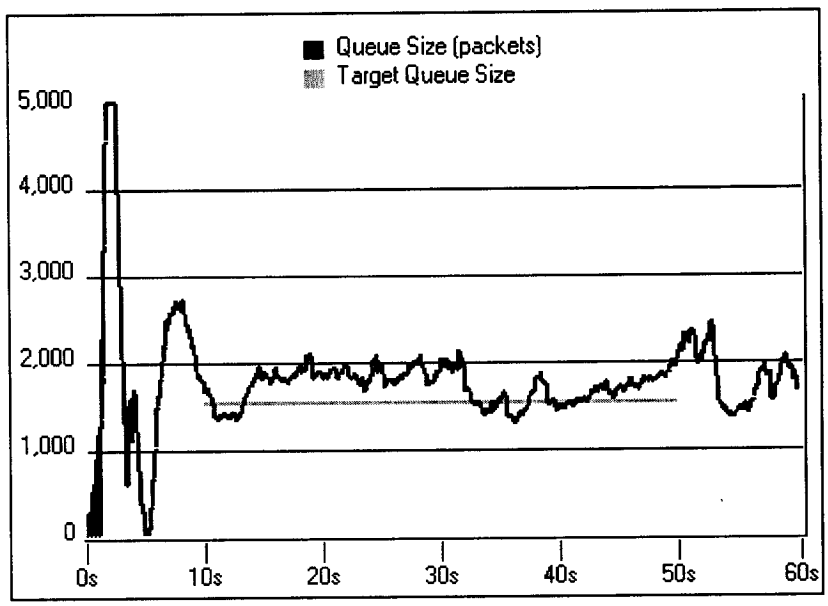

Figure 14: *Simulation 1 Results: Queue Size.*

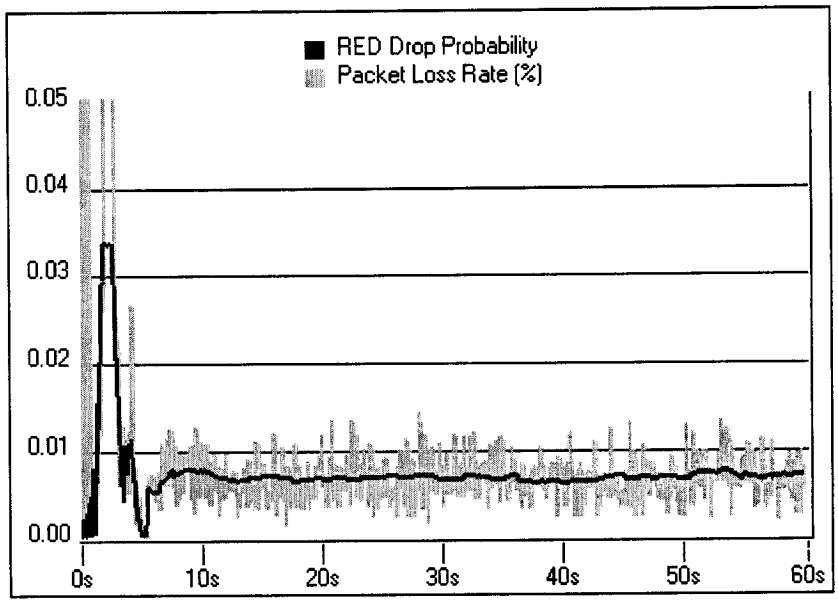

Figure 15: *Simulation 1 Results: Packet Loss Rate.*

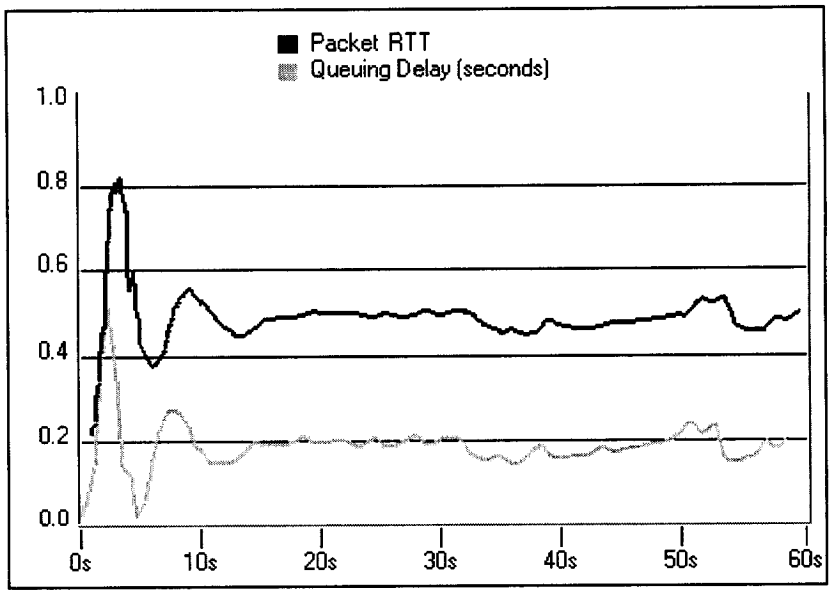

Figure 16: *Simulation 1 Results: Estimated Packet Delay.*

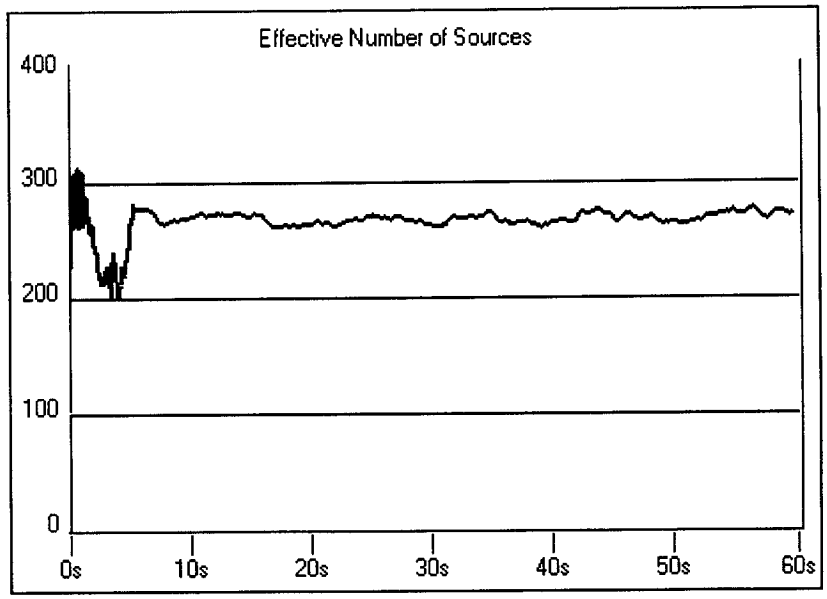

Figure 17: *Simulation 1 Results: Estimated Number of Sources.*

Figures 14 and 15 show that the queue size remains close to the target queue size for the entire simulation, and that the drop probability remains almost constant, suggesting that the system has stabilized in the manner described in the previous section. Indeed, a simple calculation verifies that the loss probability has reached its equilibrium value. Take $C = L/N = 10433/300$ and $r = T+Q/L = 0.3+1500/10433$. Then (2.13) gives

$$\frac{(1-k)^2}{k} = \left(\frac{0.4483 \cdot 34.7767}{1.3098}\right)^2, \tag{7.5}$$

which yields a solution ($k \approx 0.00696$) close to the simulated value. Note that the theory requires only the drop *probability* to be constant; the actual packet loss rate, which is based on bernoulli trials with a probability of success equal to the drop probability, may fluctuate as it is inherently random.

Figure 16 shows that the estimated round-trip time is very accurate, as it is almost exactly 300 ms more than the observed queueing delay. This shows that the round-trip time estimation method of Chapter 3 works well, even though it samples a relatively small proportion (less than 1%) of the round-trip times (although this estimate is cumulative with time).

Figure 17 shows that the RWFD algorithm is also effective at estimating the number of sources, as the observed value is close to the true number of 300. The estimate is slightly less than 300, partially due to the fact that some TCP sources are in timeout mode, and are not sending any packets.

For this example, there is no significant performance advantage over a properly tuned RED policy. However, the RWFD algorithm has automatically chosen the drop probability that will achieve the target queue size. One can check that the drop probability in Figure 15 indeed satisfies (7.4).

Example 2 is the same as Example 1, except that the number of sources is doubled to $N = 600$. The results are shown below.

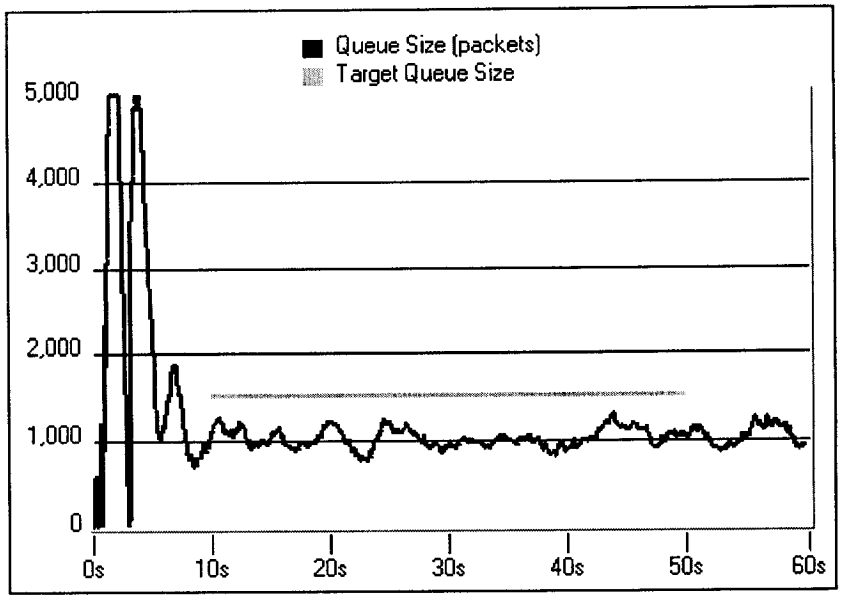

Figure 18: *Simulation 2 Results: Queue Size.*

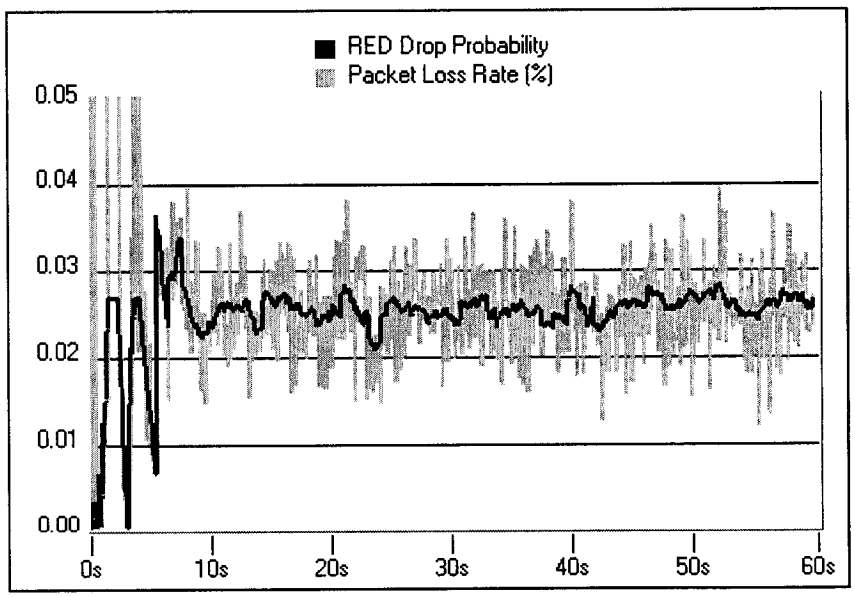

Figure 19: *Simulation 2 Results: Packet Loss Rate.*

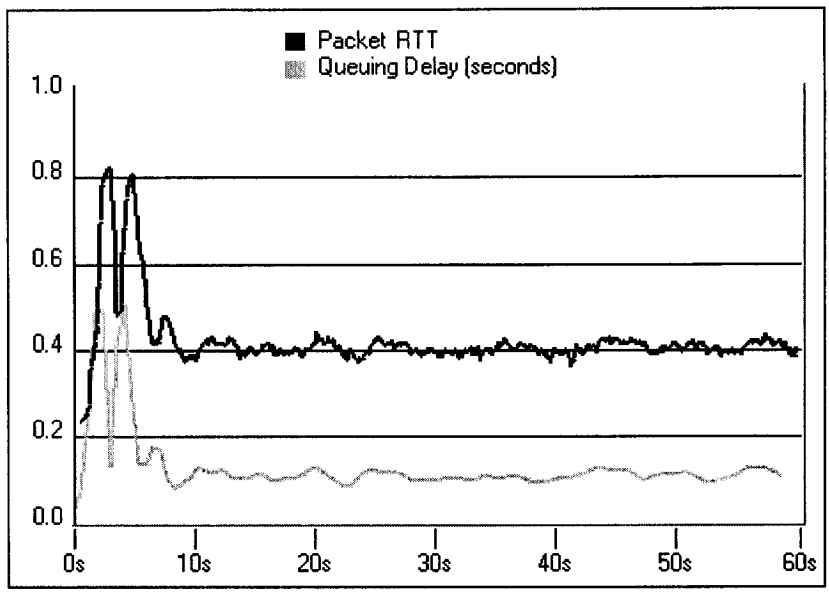

Figure 20: *Simulation 2 Results: Estimated Packet Delay.*

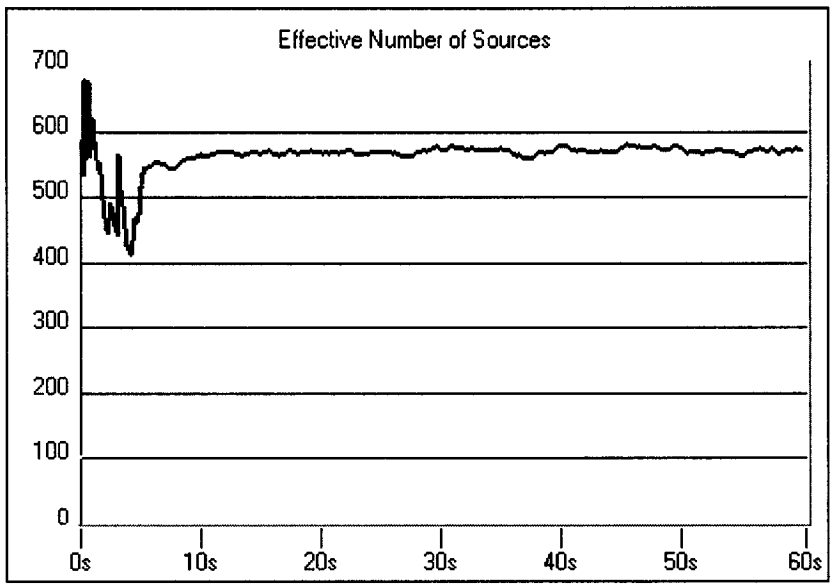

Figure 21: *Simulation 2 Results: Estimated Number of Sources.*

The round-trip time and number of sources are still well estimated in this example. However, the drop probability is not as constant as before. This can be attributed to several factors. First of all, the higher loss probability of about 2.5% corresponds to a mean window size of approximately 8 packets (as opposed to 15 packets in Example 1). The result is a larger number of TCP sources moving in and out of timeout, because they experience losses while their window size is too small. This is not accounted for in the model. Furthermore, since the RWFD algorithm must square its estimates for N, estimation errors, and the resultant noise levels, will increase nonlinearly with N. The loss probability, however, still satisfies (7.4), and it approaches the theoretical value of about 3% given by (2.13). (It is less because the loss probability is calculated based on a target queue size of 1500 packets, while the actual queue size is only 1000.)

The observation that the queue size has stabilized at a value lower than the target can also be attributed to the fact that many sources are in timeout due to the small window size.

In the final example, the network is subject to varied conditions. The transmission delays of the sources are distributed so that equal numbers of sources have external delays of 140 ms, 180 ms, 220 ms, 260 ms, and 300 ms. The (harmonic) mean external delay is then 205 ms. The number of FTP connections is changed over time, so that there are 200 sources for the first 30 seconds, 300 for 30-60 seconds, 400 for 60-90 seconds, 300 for 90-120 seconds, and 200 for 120-150 seconds. There are 450 HTTP data bursts over the 150 second simulation. The target queue size is set to 1000 packets. The results are shown below.

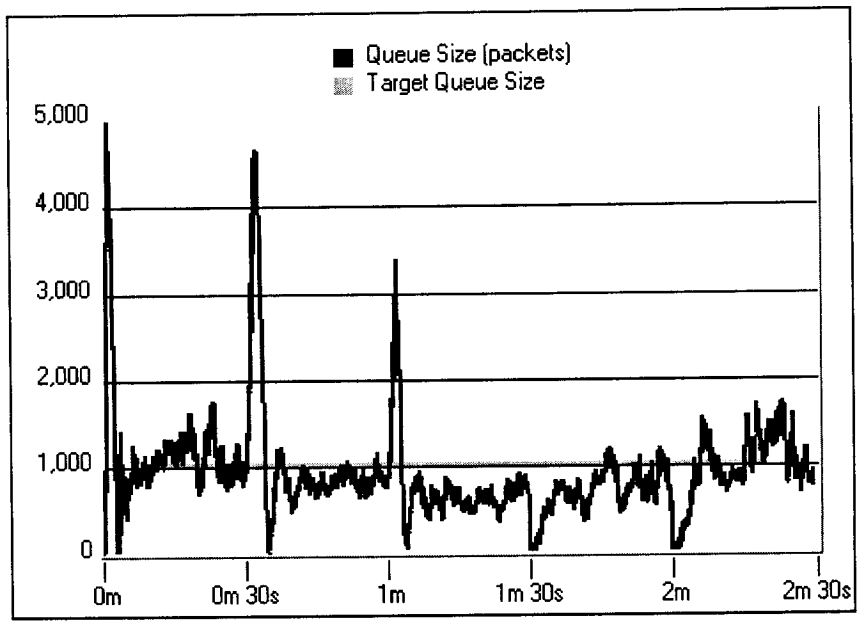

Figure 22: *Simulation 3 Results: Queue Size.*

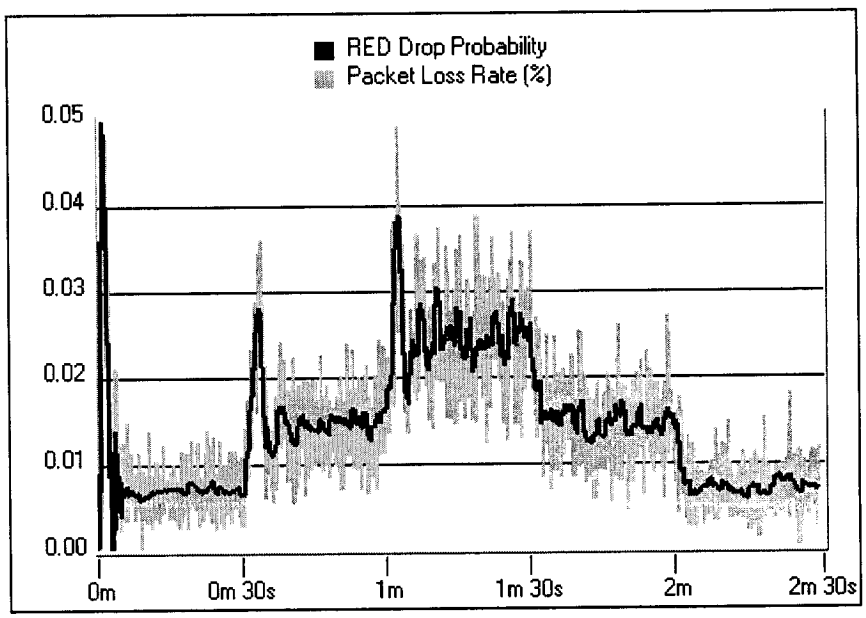

Figure 23: *Simulation 3 Results: Packet Loss Rate.*

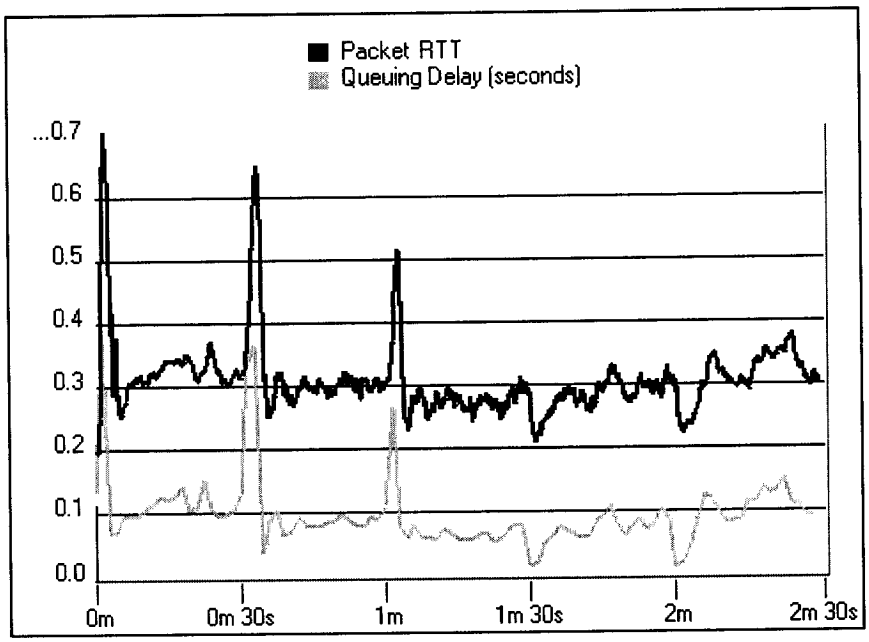

Figure 24: *Simulation 3 Results: Estimated Packet Delay.*

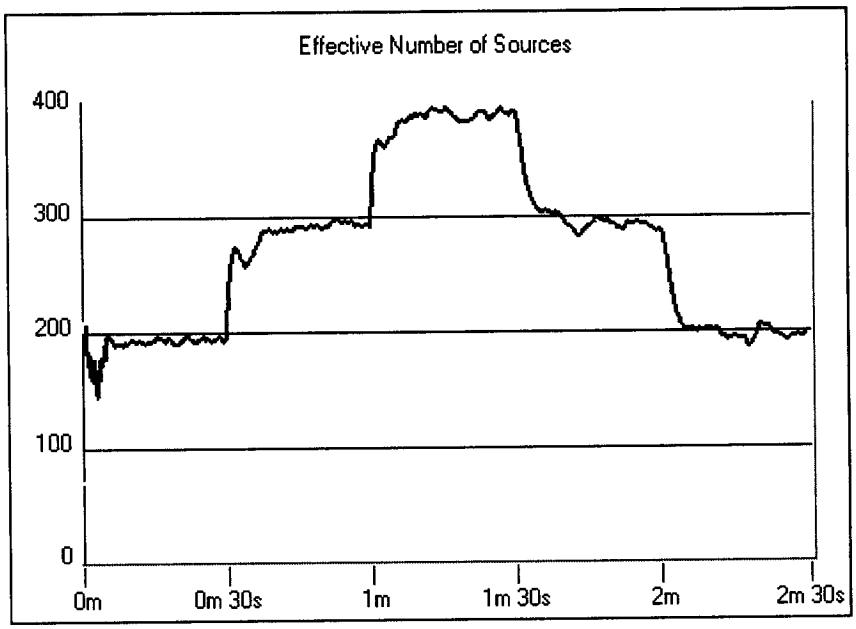

Figure 25: *Simulation 3 Results: Estimated Number of Sources.*

Again, the number of sources and the mean round-trip time are estimated quite well. The drop probability is fairly constant in each 30-second region, especially for lower numbers of sources. It adjusts with the number of sources to keep the queue close to the target value, and an easy calculation shows that the loss probability again follows the equilibrium values given by (2.13). The queue size and loss probability stabilize rapidly to the target levels after an initial transient response to each system disturbance. The algorithm is successful in adapting to varying network conditions, provided that there are reasonable periods of stability. Indeed, the settling time for this example seems to be only on the order of a few seconds.

7.4 Summary and Conclusion

It has been shown that large TCP data networks can be stabilized by randomly dropping packets with a well-chosen, constant probability, and that in practice, a nearly-constant dropping probability can be used to achieve similar results. Some knowledge of the number of sources, the transmission delay, and the transmission rate must be present in order to choose this dropping probability. Furthermore, if these network parameters are continuously monitored, the dropping probability can be periodically updated to keep the queue size close to a target level.

Network stability was guaranteed by showing that the TCP congestion window size can be modelled as a Markov chain with a unique steady-state distribution, and concluding that a large number of sources will eventually have their window sizes governed by this distribution via a mean-field argument. Furthermore, the steady-state characteristics of the network are approximately known and given by [1].

Any network with a large number of sources will thus naturally converge to a steady state, with a constant mean window size and corresponding constant queue size. The only criteria are that the network conditions, such as the number of sources, round-trip times, and loss probability, remain unchanged. The first two conditions are beyond our control, but the last is not. The conclusion is that stability is guaranteed for a static network when a loss probability giving a moderate queue size is chosen and then left unchanged as much as possible. For RED, this leads to the policy of

picking a flat or very gentle slope for the loss probability function, with a mean value that is well-tuned to the network conditions.

The RWFD algorithm satisfies this philosophy by choosing an appropriate loss probability, and by restricting its variation more and more as the network conditions become less forgiving. Although it does not impose an exactly constant loss probability, the algorithm represents a fair compromise between network stability and queue size control.

The application of general state-space Markov chain theory to the complicated issue of TCP network stability is difficult but rewarding. It is a testament to the applicability of the theory, and provides useful results to the practitioner. It is hoped that this thesis not only helps solve a practical problem, but also serve as an example of how stochastic stability theory can be applied to complex situations. The use of TCP may fall out of favor in the future, but there will always be a need to bridge the gap between theory and practice.

Bibliography

[1] BACCELLI, F., MCDONALD, D. R., REYNIER, J.(2002). A mean-field model for multiple TCP connections through a buffer implementing RED. *Performance Evaluation Vol. 11, (2002) pp. 77-97.* Elsevier Science.

[2] MEYN, S.P., TWEEDIE, R.L.(1994). Markov Chains and Stochastic Stability. Springer-Verlag. *Available online at black.csl.uiuc.edu/ meyn/pages/book.html.*

[3] ROSS, S.M.(1997). Introduction to Probability Models, 6th edition. Academic Press.

[4] BILLINGSLY, P.(1995). Probability and Measure, 3rd edition. Wiley.

[5] HOLLOT, C.V., MISRA, V., TOWSLEY, D., GONG, W-B. (2001). A control theoretic analysis of RED. *IEEE Infocom 2001.*

[6] AWEYA, J., OUELLETTE, M., DELFIN, Y. M., CHAPMAN, A. (2000). A load adaptive mechanism for buffer management. *Computer Networks, Vol. 36, Issues 5-6, August 2001, pp. 709-728.* Elsevier Science.

[7] R. T. Morris, "Scalable TCP Congestion Control", *Thesis, Harvard University, Cambridge*, Jan. 1999.

[8] FLOYD, S., JACOBSON, V. (1993). Random early detection gateways for congestion avoidance. *IEEE/ACM Trans. Networking.*, **11**, No.4 397-413.

[9] FLOYD, S. (1999). The NewReno modification to TCP's fast recovery algorithm. *Internet Engineering Task Force, RFC 2582.*

[10] FOURNIÉ, L., MCDONALD, D. R., MASKERY, M.(2003). RED Without Feedback Delay.

[11] MCDONALD, D. R., REYNIER, J.(2003). Convergence to a Mean-Field model of multiple TCP connections through a buffer implementing RED. *Preprint.*