PARTITIONING AND INVARIANCE OF AIMD DYNAMICS IN SYNCHRONISED NETWORKS

R.N.Shorten [*] **D.J.Leith** [*]

[*] *Hamilton Institute, NUI Maynooth*

Abstract: In this paper we present new results on the dynamics of networks of AIMD flows. The results reveal an invariance and partitioning property that indicates potential for the design of soundly-based adaptive AIMD strategies.

Keywords: TCP, Congestion Control, Adaptive Control

1. INTRODUCTION

A basic problem in the design of communication networks is the development of congestion control algorithms. Conventional congestion control algorithms were deployed for two principal reasons: (a) to ensure avoidance of network congestion collapse ; and (b) to control the degree of network fairness. Attempts to deal with network congestion have resulted in the widely applied Transmission Control Protocol (TCP) . While the current TCP congestion control algorithm has proved remarkably durable, it is likely to be less effective on the next generation of communication networks and it is widely agreed within the networking community that new congestion control algorithms must be developed to accompany the realization of such networks.

The task of developing such algorithms is non-trivial. In addition to the requirement of avoiding congestion collapse, fundamental requirements of congestion control algorithms include: efficient use of bandwidth; the ability to control fair allocation of bandwidth among sources; and network responsiveness. These requirements must be met while respecting key constraints including: decentralised design (TCP sources have restricted information available to them); scalability (the qualitative properties of networks employing congestion control algorithms should be independent of the size of the network and of a wide variety of network conditions); and backward compatibility with conventional TCP sources.

The balance between fairness, efficiency and responsiveness currently achieved by TCP is a delicate one that remains relatively poorly understood. Many of the proposals for changes to TCP congestion control to improve performance advocate increasing the AIMD backoff factor (e.g. High-Speed TCP, H-TCP, TCP-Westwood). However, it is known (e.g.) that network dynamics can be strongly influenced by this parameter and that changes may disrupt the existing balance between fairness, efficiency and responsiveness. In this paper we present new results on the dynamics of networks of AIMD flows, with particular emphasis on the the impact of changes to the AIMD backoff factor. We prove a surprising partitioning and invariance result that indicates the possibility of developing soundly-based adaptive AIMD schemes with arbitrary changes to the AIMD backoff factor without sacrificing network stability and responsiveness.

The paper is organised as follows. In Section 2 we begin by briefly reviewing some recent results on the modelling and analysis of TCP in networks with drop-tail queues (the prevalent queueing discipline in the internet). We then present our main results and illustrate how they might be used to form the basis of a soundly-based adaptive AIMD strategy. Finally, we conclude by summarising our results.

2. PRELIMINARIES

The TCP congestion control algorithm regulates the congestion window, *cwnd*, of a flow, which determines the number of unacknowledged packets in flight. The standard congestion control algorithm updates the congestion window *cwnd* according to an Additive Increase Multiplicative Decrease (AIMD) control law. In the congestion avoidance phase, when a source i receives a TCP ACK, it increments *cwnd* according to $cwnd \rightarrow cwnd + \alpha/cwnd$ where $\alpha = 1$ for the standard TCP algorithm. When packet loss is detected, *cwnd* is reduced by a backoff factor β: thus $cwnd \rightarrow \beta cwnd$, where $\beta = 0.5$ for standard TCP.

Following we consider communication networks for which at congestion every source experiences a packet drop. In this case the network dynamics may be modelled as follows .

$$W(k+1) = AW(k), \qquad (1)$$

where $W^T(k) = [w_1(k), \cdots, w_n(k)]$, $w_i(k)$ denotes the congestion window of flow i at the kth network congestion event and

$$A = \begin{bmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \beta_n \end{bmatrix} \qquad (2)$$

$$+ \frac{1}{\sum_{j=1}^{n} \alpha_j} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_n \end{bmatrix} \begin{bmatrix} 1-\beta_1 & 1-\beta_2 & \cdots & 1-\beta_n \end{bmatrix}.$$

The matrix A is a positive matrix (all the entries are positive real numbers) and it follows that the synchronised network (1) is a positive linear system . The following theorem will prove useful in the sequel.

Theorem 1. ; Let A be defined as in Equation (2). Then A is a column stochastic matrix with Perron eigenvector $x_p^T = [\frac{\alpha_1}{1-\beta_1}, ..., \frac{\alpha_n}{1-\beta_n}]$ and whose eigenvalues are real and positive. Further, the network converges to a unique stationary point $W_{ss} = \Theta x_p$, where Θ is a positive constant such that the constraint (**??**) is satisfied; $\lim_{k\to\infty} W(k) = W_{ss}$; convergence is geometric and the rate of convergence of the network to W_{ss} is bounded by the second largest eigenvalue of A; the second largest eigenvalue of A lies in the interval $[\beta_1, \beta_2]$ where the network backoff factors are ordered as $\beta_1 \geq \beta_2 \geq .....\beta_n$.

The following facts follow immediately from Theorem 1:

(i) **Fairness:** Congestion window fairness at each congestion event is achieved when the Perron eigenvector x_p is a scalar multiple of the vector $[1, ..., 1]$; that is, when the ratio $\frac{\alpha_i}{1-\beta_i}$ does not depend on i. Further, since it follows for conventional TCP-flows ($\alpha = 1, \beta = 1/2$) that $\alpha = 2(1-\beta)$, any new protocol operating an *AIMD* variant that satisfies $\alpha_i = 2(1-\beta_i)$ will be TCP-friendly (i.e. fair with legacy TCP flows).

(ii) **Network responsiveness:** The magnitude of the largest backoff factor β_1 bounds the convergence rate of the entire network, with the 95% rise time measured in congestion epochs bounded by $\log 0.05/\log \beta_1$. Consequently, fast convergence to the equilibrium state (the Perron eigenvector) is guaranteed if the largest backoff factor in the network is small.

3. MAIN RESULTS

We have the following main results.

3.1 Invariance and Partitioning

Let us partition the network flows into two classes (i) flows with congestion window w_i lying within a δ neighbourhood of the equilibrium value $w_i(\infty)$ i.e. with $\frac{|w_i - w_i(\infty)|}{w_i(\infty)} \leq \delta$ and (ii) other flows. Then we have the following result.

Theorem 2. Consider a network of the form described in (1). Let $w_i(0) = w_i(\infty) + \delta_i \; \delta_i \in \mathbb{R}$ denote the initial condition of the i'th flow $w_i(k)$, and $w_i(\infty)$ denote the asymptotic value. Let the network backoff factors be ordered according to $\beta_1 \geq \beta_2 \geq .. \geq \beta_n$, with $\beta_i = \underline{\beta}$ for all $i \in [m+1, n]$. Suppose that $\| \frac{\underline{\beta} - \beta_j}{1-\beta_j} \| \leq 1$ for all $j \in [1, m]$. Then the following statements are true.

(P1) Invariance.
Let $\delta = sup\{\frac{|w_1(0) - w_1(\infty)|}{w_1(\infty)}, ..., \frac{|w_m(0) - w_m(\infty)|}{w_m(\infty)}\}$. Then $\frac{|w_i(k) - w_i(\infty)|}{w_i(\infty)} \leq \delta$ for all $k > 0$ and for $i \in [1, m]$.

(P2) Convergence.
$\frac{|w_i(k) - w_i(\infty)|}{w_i(\infty)} \leq |\underline{\beta}^k \delta_i| + |\delta|$ for all $k > 0$, $i \in [m+1, n]$.

Proof 1. We prove each of our claims in turn. For convenience we assume without any loss of generality that $w_i(\infty) = \frac{\alpha_i}{1-\beta_i}$.

Property P1 : Denote the i'th component of the vector $AW(0)$ by v_i where $1 \leq i \leq m$. We have:

$$\frac{v_i - w_i(\infty)}{w(\infty)} = \beta_i \delta_+ (1-\beta_i) \sum_{j=1}^{m} \frac{\underline{\beta} - \beta_j}{1-\beta_j} \alpha_j \delta_j$$

$$= \beta_i \delta_i + (1 - \beta_i)\Delta$$

Since $\beta_i \in [0, 1]$ it follows from convexity that

$$|\frac{v_i - w_i(\infty)}{w(\infty)}| \leq max\{|\delta_i|, |\Delta|\},$$

for all $1 \leq i \leq m$. Now assume that $\underline{\beta}$ and β_j are chosen such that $\| \frac{\underline{\beta}-\beta_j}{1-\beta_j} \| \leq 1$ for all $1 \leq j \leq m$. One can see that this is always possible since $\underline{\beta}$ is fixed, and the other β_j are upper bounded by β_1. Then

$$|\sum_{j=1}^{m} \frac{\underline{\beta} - \beta_j}{1 - \beta_j} \alpha_j \delta_j| \leq max\{|\delta_j|\}, \qquad (3)$$

for all $1 \leq j \leq m$. Hence, it follows that

$$|\frac{v_i - w_i(\infty)}{w(\infty)}| \leq |\delta|,$$

for all $1 \leq i \leq m$.

Property P2 : In the spirit of $P1$ we have that

$$|\frac{w_i(k) - w_i(\infty)}{w(\infty)}| \leq |\underline{\beta}^k \delta_i| + (1 - \underline{\beta}^k)|\delta|,$$

for all $m + 1 \leq i \leq n$ and for all $k > 0$. It follows that

$$|\frac{w_i(k) - w_i(\infty)}{w(\infty)}| \leq |\underline{\beta}^k \delta_i| + |\delta|,$$

as claimed.

Theorem 2 states (i) that flows which start within a δ neighbourhood of equilibrium will remain within a δ neighbourhood for all time i.e. perturbations to other network flows have a strictly limited impact on flows already close to equilibrium, and (ii) flows starting from outside a δ neighbourhood of equilibrium will converge geometrically to a δ neighbourhood at rate $\underline{\beta}$ i.e. the convergence rate is unaffected by the backoff factors of those flows already close to equilibrium.

3.2 Convergence of Time-Varying AIMD Networks

Suppose the flow AIMD parameters α and β are now time-varying so that the network dynamics become

$$W(k + 1) = A(k)W(k) \qquad (4)$$

where $A(k)$ is appropriately defined. We have the following convergence result for this time-varying systems.

We begin formally by denoting for any $T \in \mathbb{R}^{n \times n}$, T^* as the restriction of T to $n - 1$ dimensional subspace S that is orthogonal to vector $y^T = [1, 1 \ldots, 1]$. Recall that if T is column stochastic and positive then T^* is contraction on S and therefore $\|T^*\| < 1$.

Theorem 3. Let $\{A(k)\}_{k \in \mathbb{N}}$ be a stochastic process which corresponds to synchronized, time-varying, network (this means that $A(k)$ is a positive matrix and is am element of a finite set of positive matrices $\mathcal{M}$). If all matrices in $\mathcal{M}$ have common right Perron eigenvector x_p then $W(k)$ converges to $x_p y^T W(0)$. Moreover convergence is geometrical with convergence rate not larger then

$$\mu = max\{\|M^*\| : \ M \in \mathcal{M}\}$$

Proof 2. The proof follows from standard results on infinite products of positive matrices.

Theorem 3 states that the network (4) will converge to a unique fixed point if each of the matrices in the set $\mathcal{M}$ has the same Perron eigenvector. Consider the family $\Sigma(A), A \in \mathcal{M}$ of time-invariant systems $\Sigma(A) : W(k+1) = AW(k)$. The equilibrium point of $\Sigma(A)$ is determined by the Perron eigenvector of A. The requirement that the A share the same Perron eigenvector is equivalent to the requirement that the $\Sigma(A)$ share the same equilibrium point.

4. APPLICATION OF MAIN RESULTS: ADAPTIVE AIMD

The TCP congestion control has undergone a number of changes since the original design by Jacobson and this process continues as the internet evolves. For example, increasing link speeds have led to paths with high bandwidth-delay product becoming more common and it known that the current TCP congestion control algorithm can exhibit very sluggish convergence on such paths. This occurs because while the network may have a 95% rise time bounded by 4 congestion epochs (see above), the *duration* of each congestion epoch scales linearly with bandwidth-delay product. Link speeds have increased by several orders of magnitude over the last decade or so, with consequent impact on the TCP congestion epoch durations experienced. Many of the proposals for changes to TCP congestion control to improve performance on high bandwidth-delay paths advocate increasing the AIMD backoff factor (e.g. High-Speed TCP, H-TCP, TCP-Westwood). Larger backoff factors mean that flows reduce their congestion window by less upon detecting network congestion and thus are less likely to empty network queues. As a result, link utilisation tends to increase as the backoff factor increased. Increasing the backoff factor will, unfortunately, have a negative impact on network convergence times unless corrective action is taken, see for example Figure 1.

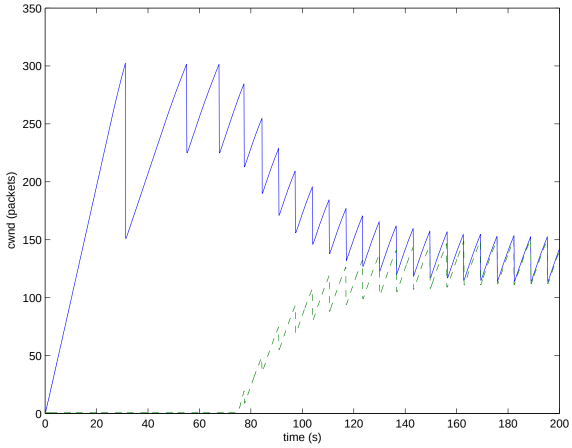

Fig. 1. Illustrating poor responsiveness with larger backoff factors using an NS network simulation of a network with a 20Mb bottleneck link, a 150ms delay, a maximum queue size of 50 packets and backoff factor of 0.75.

The results in the present paper suggest, however, that we can design a number of controllers to address the different performance requirements (a controller to ensure high network utilisation, a controller to ensure rapid convergence) and switch between these as network conditions change[1]. Adaptive algorithms that involve mode switching are known to be difficult to design and to analyse .

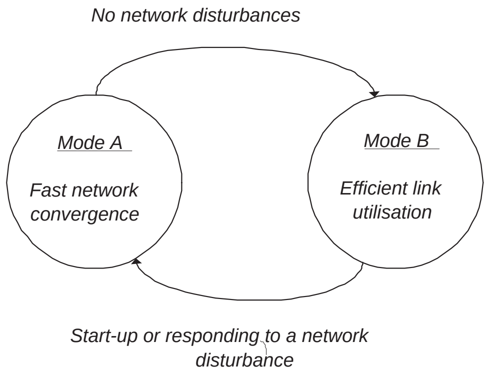
No network disturbances
Mode A
Fast network convergence
Mode B
Efficient link utilisation
Start-up or responding to a network disturbance

Fig. 2. Switched adaptive congestion control algorithm.

To see how we might use Theorem 2 in this context, let us partition the network flows into two classes (i) flows with congestion window w_i lying within a δ neighbourhood of the equilibrium value $w_i(\infty)$ i.e. with $\frac{|w_i - w_i(\infty)|}{w_i(\infty)} \leq \delta$ and (ii) other flows. Select the backoff factors of class (i) flows to be $\beta_i \leq \bar{\beta}$. Select the backoff factors of class (ii) flows to be a small value $\underline{\beta}$ such that $|\frac{\underline{\beta}-\overline{\beta}}{1-\overline{\beta}}| \leq 1$. Then we have from property (P1) that the class (i) flows will remain within a δ neighbourhood of the equilibrium for all time while from property (P2) we have that class (ii) flows will converge geometrically to a δ neighbourhood of the equilibrium at rate $\underline{\beta}$. Notice that the latter convergence rate is determined solely by the backoff factor $\underline{\beta}$ of the class (ii) flows; choosing $\underline{\beta} = 0.5$ we recover the convergence rate of standard TCP. In other words, we may select a $\bar{\beta}$ and $\underline{\beta}$ (largest and smallest backoff factor) and specify a disturbance threshold δ. Any source whose perturbation from the equilibrium is within this threshold will remain in this bounded region and can choose their backoff factors freely (in particular, they can choose their backoff factors to maximise link utilisation). Sources that are perturbed outside of this region can ensure rapid convergence to the region as k increases by selecting a small backoff factor.

This yields the following decentralised switching strategy,

$$\beta_i(k+1) = \begin{cases} \beta_i & |\delta_i| \leq \delta \\ \underline{\beta} & \text{otherwise.} \end{cases} \tag{5}$$

where $\beta_i = min[\frac{RTT_{min,i}}{RTT_{max,i}}, \overline{\beta}]$, $\delta_i = \frac{|w_i - w_i(\infty)|}{w_i(\infty)}$ and $\underline{\beta}$, $\overline{\beta}$ are selected to satisfy $|\frac{\underline{\beta}-\overline{\beta}}{1-\overline{\beta}}| \leq 1$.

Of course such an adaptive strategy yields time-varying network dynamics and immediately raises questions as to its impact on network stability properties. Theorem 3, however, guarantees stability under mild conditions. Specifically, it follows from the previous discussion that the $A \in \mathcal{M}$ matrices share a common Perron eigenvector $[x_1 ... x_n]^T$ when

$$\frac{\alpha_i(A)}{1-\beta_i(A)} = x_i \tag{6}$$

where $\alpha_i(A), \beta_i(A)$ denote the AIMD parameters used in matrix $A \in \mathcal{M}$ for the ith flow. Equation (6) states that to ensure a common Perron eigenvector we require that variations in the AIMD parameters α_i, β_i of the ith flow be constrained such that the ratio $\alpha_i/(1-\beta_i)$ remains constant. Observe that we have that this ratio is 2 for standard TCP. We therefore have immediately that constraining the ratio $\alpha_i(A)/(1-\beta_i(A))$ to have the value 2 for all flows simultaneously yields (i) a common Perron eigenvector that ensures a unique fixed point that is fair and (ii) ensures backward compatibility and TCP friendliness.

[1] Note that TCP currently includes a slow start mode to accelerate convergence at startup of a new flow. However, slow start action is essentially confined to the first congestion epoch following flow startup and cannot be invoked to deal with the effect of network disturbances during the 'congestion avoidance' phase of TCP.

The complete switched adaptive AIMD algorithm is therefore

$$\beta_i(k+1) = \begin{cases} \beta_i & |\delta_i| \leq \delta \\ \underline{\beta} & \text{otherwise.} \end{cases} \tag{7}$$

$$\alpha_i(k+1) = 2(1-\beta_i(k+1)) \tag{8}$$

where $\beta_i = min[\frac{RTT_{min,i}}{RTT_{max,i}}, \overline{\beta}]$, $\delta_i = \frac{|w_i - w_i(\infty)|}{w_i(\infty)}$ and $\underline{\beta}$, $\overline{\beta}$ are selected to satisfy $|\frac{\underline{\beta}-\overline{\beta}}{1-\overline{\beta}}| \leq 1$.

Comment 1: Typically we might use $\underline{\beta} = 0.5$ and $\overline{\beta} = 0.75$.

Comment 2: The switching threshold δ is a design parameter that determines the performance trade-off between efficiency and responsiveness. When δ is large, the mode switch is rarely invoked and the switching strategy reduces to the previously discussed adaptive backoff strategy. When $\delta = 0$, the switching strategy corresponds to the standard TCP strategy with backoff factor $\underline{\beta}$.

Comment 3: This decentralised switching strategy requires that the ith flow can measure or infer distance from the equilibrium congestion window $w_i(\infty)$. In current networks $w_i(\infty)$ generally cannot be measured and so this distance must be estimated. There are many ways in which such estimation might be carried out. One simple approach is to estimate the distance from the magnitude of $w_i(k+1) - w_i(k)$.

4.1 Example

The impact on responsiveness of introducing a mode switch is illustrated in Figure 3. The network conditions are identical to those in Figure 1, with a backoff factor β of 0.75 for efficient link utilisation. The additive increase parameter α is adjusted with β such that $\alpha/(1-\beta) = 2$ and so α reduces when β increases, as is evident in the figure. When a second flow starts at 75s, each flow decreases its backoff factor to 0.5, reverting to 0.75 when the congestion window is within 10% of its equilibrium value. The 95% rise time to this boundary region is 4 congestion epochs owing to the 0.5 backoff factor used, compared to 11 congestion epochs when the backoff factor is 0.75. Note that while in this example both flows switch to small backoff factors in response to the change in network conditions, in general only those flows that are perturbed outside the boundary region need adjust their backoff factors to ensure fast convergence. This is illustrated, for example, in Figure 4 where a network of 10 flows subject to a cross-flow disturbance between 200s and 205s is considered. After the cross-flow disturbance ends at 205s, it can be seen that the network rapidly converges back to equilibrium. In this example only 5 out of the 10 flows move outside the boundary region and reduce their backoff factors to 0.5.

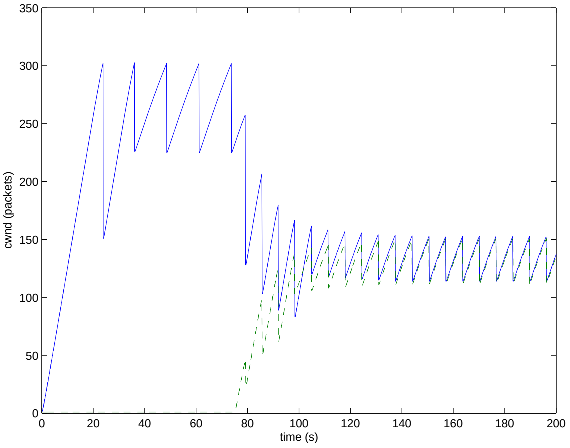

Fig. 3. Improvement in responsiveness with adaptive mode switch compared with Figure 1. (NS simulation, 20Mb bottleneck link, 150ms delay, queue size 50 packets).

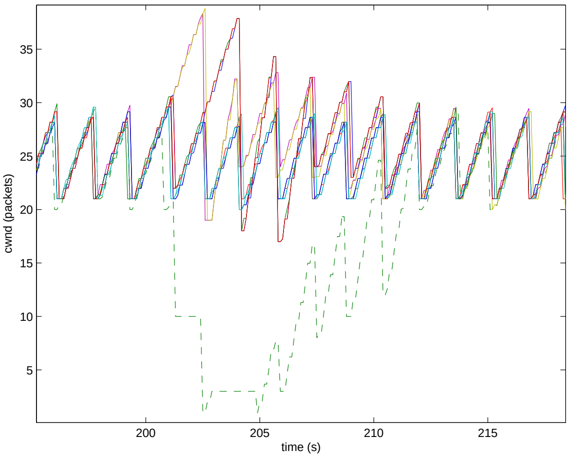

Fig. 4. Adaptive mode switch for multiple flows subject to a cross-flow disturbance from 200s to 205s. (NS simulation, 10 flows, 20Mb bottleneck link, 150ms delay, queue size 50 packets).

5. CONCLUDING REMARKS

In this paper we present new results on the dynamics of networks of AIMD flows. The results reveal a surprising invariance and partitioning property that suggests the potential exists for the design of soundly-based adaptive AIMD strategies. We show that adaptation can be used to achieve high network utilisation without compromising rapid network convergence.

We also note that the results in this paper are focussed on networks that exhibit drop synchronisation. While relatively few networks exhibit this property, we argue that the study of such networks is a useful starting point for developing analysis tools suited to drop-tail environments. Further, even though drop synchronisation is relatively rare in real networks, it is a fact that some networks do experience synchronisation (for example, in some long distance networks), and consequently such networks merit study in their own right. In this context we note several previous studies in this area ; ; ; ; .

6. ACKNOWLEDGEMENTS

This work was supported by Science Foundation Ireland grants 00/PI.1/C067 and IN3/03/I346.

REFERENCES

V. Jacobson, "Congestion avoidance and control," in *Proceedings of SIGCOMM*, 1988.

S. Floyd and K. Fall, "Promoting the use of end-to-end congestion control in the internet," *IEEE/ACM Transactions on Networking*, vol. 7, no. 4, pp. 458–472, 1999.

Various authors, "Special issue on TCP performance in future networking environments," *IEEE Commuications magazine*, vol. 39, no. 4, pp. –, 2001.

S. Floyd, "High speed TCP for large congestion windows," tech. rep., Internet draft draft-floyd-tcp-highspeed-02.txt, work in progres, February 2003.

D. Leith and R. Shorten, "H-TCP protocol for high-speed long-distance networks. proc. 2nd workshop on protocols for fast long distance networks." Proc. of Workshop on Protocols for Fast Long-Distance Networks, 2004.

L. Massoulie, "Stability of distributed congestion control with heterogeneous feedback delays," *IEEE Transactions on Automatic Control*, vol. 47, no. 6, pp. 895–902, 2002.

R. Shorten, D. Leith, J. Foy, and R. Kilduff, "Analysis and design of synchronised communication networks." Accepted for publication by Automatica, 2004.

A. Berman and R. Plemmons, *Nonnegative matrices in the mathematical sciences.* SIAM, 1979.

A. Berman, R. Shorten, and D. Leith, "Positive matrices associated with synchronised communication networks." Accepted for publication in Linear Algebra and its Applications, 2003.

D. Hartfiel, *Nonhomogeneous matrix products.* World Scientific, 2002.

S. Morse, *Control Using Logic Based Switching.* Springer Verlag, 1997.

L. Xu, K. Harfoush, and I. Rhee, "Binary increase congestion control for fast long-distance networks." To appear in Proceedings of IEEE INFOCOM, 2004.

D. Hong and D. Lebedev, "Many TCP user asymptotic analysis of the AIMD model," Tech. Rep. INRIA Technical Report 4229, INRIA Rocquencourt, 2001.

F. Bacelli and D. Hong, "Interaction of TCP flows as billiards," Tech. Rep. INRIA Technical Report 4437, INRIA Rocquencourt, 2002.

P. Brown, "Resource sharing of TCP connections with different round trip times," in *Proceedings of IEEE INFOCOM*, (Tel Aviv, Israel), March 2000.

E. Altman, T. Jimenez, and R. Nunez-Queija, "Analysis of two competing TCP/IP connections," *Perform. Evaluation*, vol. 49, no. 1-4, pp. 43–55, 2002.

J. Hespanha, S. Hohacek, K. Obrarzka, and J. Lee, "Hybrid model of TCP congestion control," in *Hybrid Systems: Computation and Control*, pp. 291–304, 2001.