TCP/IP Interaction Based on Congestion Price: Stability and Optimality

Jiayue He
Electrical Engineering
Princeton University
Email: jhe@princeton.edu

Mung Chiang
Electrical Engineering
Princeton University
Email: chiangm@princeton.edu

Jennifer Rexford
Computer Science
Princeton University
Email: jrex@cs.princeton.edu

***Abstract*— Despite the large body of work studying congestion control and adaptive routing in isolation, much less attention has been paid to whether these two resource-allocation mechanisms work well *together* to optimize user performance. Most analysis of congestion control assumes static routing, and most studies of adaptive routing assume that the offered traffic is fixed. In this paper, we analyze the *interaction* between congestion control and adaptive routing, and study the stability and optimality of the joint system. Previous work has shown that the system can be modelled as a joint optimization problem that naturally leads to a primal-dual algorithm with shortest-path routing using congestion prices as the link weights. In practice, the algorithm is commonly unstable. We consider three alternative timescale separations and examine the stability and optimality of each system. Our analytic characterizations and simulation experiments demonstrate how the step size of the congestion-control algorithm affects the stability of the system, and how the timescale of each control loop and homogeneity of link capacities affect system stability and optimality. The stringent conditions imposed for stability suggests that congestion price would be a poor feedback mechanism in practice.**

Keywords: Network utility maximization, Congestion control, Dynamic routing, TCP/IP.

I. Introduction

There are two main ways in the Internet to adapt the allocation of network resources to maximize user utility: congestion control (in TCP) and routing (in IP). Congestion control allocates the limited capacity on each link to competing flows, while routing determines which flows pass through which links. Optimization frameworks have provided rigorous characterizations of TCP and IP performance in isolation. For example, recent work has shown that TCP congestion control implicitly solves network-utility maximization problems [1], [2], [3], [4], [5], but these studies assume a static mapping of traffic to network paths. Similarly, research on traffic engineering [6], [7] and load-sensitive routing [8], [9], [10] investigate how to optimize the assignment of traffic to paths, but assume that the sources do not adapt their sending rates to the prevailing network conditions. In practice, however, these two resource-allocation mechanisms *do* interact with each other in potentially complicated ways.

Optimization-theoretic analysis of the TCP/IP interaction is scarce in the literature. For example, [11] examines the interaction between congestion control and adaptive routing based on centrally minimizing the maximum link utilization. However, congestion control is not modelled analytically and the results are limited to networks with a single bottleneck. The only paper with a detailed analytic model of TCP/IP interaction is the recent work [12]. This study views the joint optimization problem as maximizing user utility with both the source rates and network paths as optimization variables. In particular, the interaction between TCP and IP is modelled as dynamic routing based on congestion prices on the links, where congestion price can be interpreted as link metrics like packet loss or queuing delay. The work in [12], however, assumes a particular separation of timescales: congestion control converges instantaneously, followed by one step of dynamic route optimization, and the process repeats. In reality, the joint system consists of two distributed control loops running concurrently, with timescales determined by many complex factors (e.g., round-trip time, TCP session duration, routing-protocol timer, and traffic-engineering practice).

In this paper, we present a comprehensive framework to study TCP/IP interaction based on congestion price, and examine the following key questions through both analysis and simulation. **Stability**: Does the TCP/IP system converge? **Optimality**: If the system converges, does it converge to a joint optimum? Further, what kind of general conclusions can we draw to guide the design and operation of IP networks?

Given that neither shortest-path routing nor network-utility maximization has closed-form solutions in general, the analytic results in this paper focuses on stability conditions for a ring topology. Simulation results are used to further quantify the intuitions on optimality gap, general topologies, and guidelines for system improvement. We study three different timescale separations between congestion control and routing, all motivated by Internet reality:

1) **System Model One** imitates traffic engineering today where the operator tunes link weights, [6], [7]. In this model, congestion control would iterate until convergence to produce source rate and congestion price, then routing would also iterate until convergence to produce a new routing, and so on.
2) **System Model Two** is motivated by load-sensitive dynamic routing assuming that congestion control adapts at a much smaller timescale than routing, covering the model in [12] as a special case. In this model, congestion control iterates until convergence to produce source rates

and congestion prices, but routing iterates only once.

3) **System Model Three** is also motivated by load-sensitive routing but assumes that congestion control and routing adapt on the same timescale. In this model, congestion control and routing interact completely dynamically, each iterating once with overlap in their control loops.

For the ring topology, we find the conditions for all three systems to converge to minimum-hop routing. System Model One requires the initial routing configuration to be minimum-hop routing (Theorem 1) for convergence. By choosing a small enough step size (representing the steepness of the congestion-control adaptation), convergence can be guaranteed (Theorem 2) for System Model Two. In addition to a small step size, a certain capacity distribution is also required for System Model Three to converge (Theorem 3). System Model One and Two can only converge to shortest-hop routing, which may be suboptimal. Unlike the other two systems, System Model Three may also converge to other routing configurations (Theorem 3). Simulation shows that, when System Model Three converges, it is close to the optimum.

From our analysis and simulation experiments, we observe that congestion price is not an appropriate "layering price" for TCP/IP interaction, given the stringent stability condition. The following new insights are also obtained:

1) **'Small step size improves system convergence'**: The classical tradeoff between convergence (small step size helps) and speed of convergence (large step size helps) for congestion control carries over to TCP/IP interaction.
2) **'Shorter timescale enhances optimality'**: The more dynamic the interaction between congestion control and routing, the smaller the suboptimality gap between the convergent point and the jointly optimal TCP/IP solution, because information passed between TCP and IP is less stale.
3) **'Homogeneity enhances optimality'**: Optimality of TCP/IP interactions are enhanced by 'homogeneity' of link capacities.

The rest of the paper is organized as follows. Section II provides the network topology, routing, and congestion control models, followed by Section III that describes the details of three System Models. Analysis and simulation are presented in Sections IV and V, respectively. Future research directions are then outlined in the conclusion section VI.

II. Models and Notation

We start with some general assumptions. First, we will only consider a single Autonomous System, so that shortest-path (minimum-cost) routing based on link weights (link costs) is a reasonable model. Second, we will consider a routing model where traffic between source-destination pairs can be split arbitrarily between multiple paths. This is not the OSPF [13] or IS-IS protocols used today, but can be easily implemented using the emerging MPLS [14] technology. Thirdly, we assume the sources have infinite backlog.

The notation follows [12]: in general, small letters are used to denote vectors, e.g., x with x_i as its ith component; capital letters to denote matrices, e.g., H, W, R, or constants, e.g., L, N, K^i; and script letters to denote sets of vectors or matrices, e.g., $\mathcal{W}_m$, $\mathcal{R}_m$. Superscript is used to denote vectors, matrices, or constants pertaining to source i, e.g., w^i, H^i, K^i.

A. Network and Routing

A network is modeled as a set of L bi-directional links with finite capacities $c = (c_l, l = 1, \ldots, L)$, shared by a set of N source-destination pairs, indexed by i (we will also refer to a source-destination pair simply as "source i"). There are a total of K^i acyclic paths for each source i, represented by a $L \times K^i$ 0-1 matrix H^i, where

$$H^i_{lj} \quad = \quad \begin{cases} 1, & \text{if path } j \text{ of source } i \text{ uses link } l \\ 0, & \text{otherwise.} \end{cases}$$

Let $\mathcal{H}^i$ be the set of all columns of H^i that represents all the available paths for i. Define the $L \times K$ matrix H as

$$H \quad = \quad [H^1 \ \ldots \ H^N],$$

where $K := \sum_i K^i$. H defines the topology of the network.

Let w^i be a $K^i \times 1$ vector where the jth entry represents the fraction of i's flow on its jth path such that

$$w^i_j \geq 0 \ \ \forall j, \quad \text{and} \quad \mathbf{1}^T w^i = 1,$$

where $\mathbf{1}$ is a vector of an appropriate dimension with the value 1 in every entry. We allow $w^i_j \in [0, 1]$ for multipath routing. Collect the vectors w^i, $i = 1, \ldots, N$, into a $K \times N$ block-diagonal matrix W. Define the corresponding set $\mathcal{W}_m$ for multipath routing as $\{W |\ W = \text{diag}(w^1, \ldots, w^N) \in [0, 1]^{K \times N},\ \mathbf{1}^T w^i = 1\ \}$.

In summary, H defines the set of acyclic paths available to each source, and represents the network topology. W defines how the sources load balance across these paths. Their product defines a $L \times N$ routing matrix $R = HW$ that specifies the fraction of source i's flow that traverses each link l.

B. Review TCP Model

As in [4], we interpret the equilibria of various TCP congestion-control algorithms as solutions of a network utility maximization problem defined in [1], [**?**]. Suppose each source i has a utility function $U_i(x_i)$ as a function of its total transmission rate x_i. We assume U_i is increasing and strictly concave (as is the case for TCP algorithms [4]). The constrained utility maximization problem over x for a fixed R is

$$\begin{array}{ll} \text{maximize} & \sum_i U_i(x_i) \\ \text{subject to} & Rx \ \leq \ c. \end{array} \tag{1}$$

The duality gap for the above optimization problem is zero. Zero duality gap means that the minimized objective value of the Lagrange dual problem is equal to the maximized total utility in the primal problem (1).

We briefly review the solution to (1). First form the Lagrangian of (1):

$$L(x, p) = \sum_i U_i(x_i) + \sum_l p_l \left(c_l - y_l\right)$$

where $p_l \geq 0$ is the Lagrange multiplier (*i.e.*, congestion price) associated with the linear flow constraint on link l, and $y_l = \sum_i R_{li} x_i$ is the load on link l. It is important that the Lagrangian can be decomposed for each source:

$$\begin{aligned} L(x,p) &= \sum_i \left[U_i(x_i) - \left(\sum_l R_{li} p_l \right) x_i \right] + \sum_l c_l p_l \\ &= \sum_i L_i(x_i, q_i) + \sum_l c_l p_l \end{aligned}$$

where $q_i = \sum_l R_{li} p_l$ is the end-to-end price for source i.

The Lagrange dual function $g(p)$ is defined as the maximized $L(x,p)$ over x for a given p. This 'net utility' maximization can be conducted distributively by each source, as long as the aggregate link price q_i is feedback to source i:

$$x_i^*(q_i) = \underset{x_i}{\text{argmax}} \; [U_i(x_i) - q_i x_i], \quad \forall i. \tag{2}$$

The Lagrange dual problem of (1) is to minimize $g(p)$ over $p \geq 0$. An iterative gradient method can be used to update the dual variables p in parallel on each link to solve the dual problem:

$$p_l(t+1) = \left[p_l(t) - \alpha \left(c_l - \sum_i R_{li} x_i^*(q_i(t)) \right) \right]^+, \quad \forall l \tag{3}$$

where t is the iteration number and $\alpha > 0$ is step size. It can be shown [4] that, for sufficiently small step size, the above updates of (x,p) through (2,3) converge to the jointly optimal rate allocation and congestion prices for (1) and its Lagrange dual problem. At equilibrium, the following Karush-Kuhn-Tucker (KKT) optimality conditions [4] are satisfied:

$$\begin{aligned} & q_i = U_i'(x_i) \quad \forall i \\ & y_l \begin{cases} \leq c_l & \text{if } p_l = 0 \\ = c_l & \text{if } p_l > 0 \end{cases} \quad \forall l \\ & x \geq 0, \quad p \geq 0. \end{aligned} \tag{4}$$

III. Problem Formulations

We start the investigation by considering the joint TCP/IP optimization problem and motivate the usage of congestion price. Then we define three models, comparing and contrasting their timescale assumptions. We conclude this section by motivating the usage of a ring topology for our analysis and some of the simulation experiments.

A. Joint Optimization Model

What kind of TCP/IP interactions would work together to maximize end-user utilities over both rate allocation x and routing matrix R, solving the following problem:

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_i U_i(x_i) \\ \text{subject to} \quad & Rx \leq c, \; x \geq 0 \\ & R \in \mathcal{R}, \end{aligned} \tag{5}$$

where both R and x are both variables?

Consider the dual problem of (5) in the form of optimizing the Lagrangian $L(p,x,R)$:

$$\min_{p \geq 0} \sum_i \max_{x_i \geq 0} \left(U_i(x_i) - x_i \min_{R \in \mathcal{R}} \sum_l R_{li} p_l \right) + \sum_l c_l p_l. \tag{6}$$

It hints that dynamic shortest-path routing $\min_R \sum_l R_{li} p_l$, where link cost is based on congestion prices p, may be designed to jointly maximize network utility with TCP. This possibility was first investigated in [12], which shows that, under a particular timescale separation, TCP/IP would jointly solve (5) if an equilibrium exists. Such an equilibrium exists if multipath routing is allowed, but it can be unstable. It can be stabilized by adding a static component to link weight, but at the expense of a reduced utility at equilibrium.

Before giving the detailed description of the models, we highlight the following basic intuition: TCP adjusts x, IP adjusts R, each affected by the other through the congestion-price vector $p(x,R)$, which is clearly a function of both x and R, and jointly determining the objective of $\sum_i U_i(x_i)$. Since the timescale of TCP is affected by the round-trip time and that of IP determined by routing protocols and operational practice, there can be four different models of the above interaction. Given that IP rarely operates faster than TCP convergence, we have three System Models, including the one in [12] as a special case, described below.

B. System Model Definitions

The progression offered by Figures 1, 2 and 3 shows a trend toward a tighter coupling of the two control loops:

1) System Model One: The TCP loop shows the steps taken for congestion control as described in Section II.B, and, given (x^*, p^*) from TCP, the IP loop is as follows: (i) update the congestion price p_l for link l, given the link load y_l, (ii) update the routing per source given the link weights set to the congestion prices p_l, and (iii) update the link loads y_l based on the new routing matrix R. Then the TCP loop is repeated, followed by another round of the IP loop, and so on.
2) System Model Two: TCP is exactly the same as in Model One, but the IP loop iterates only once. This is similar timescale separation proposed by [12]. However, each round of IP model in [12] ignores the change in link load y_l due to change in routing (and can also be viewed as setting the step size to zero). Each IP round in our Model Two takes a full iteration of an IP round in Model One by taking into account the effect on link load due to the anticipated routing change.
3) System Model Three: The TCP and IP loop are interacting at the same timescale. Each TCP/IP round consists of maximization over x and minimization over R of the Lagrangian (6) for the same given p, which then is updated based on both the change in x and that in R.

C. Ring Topology and Traffic Model

One of the goals of this paper is to derive closed-form solutions for the stability conditions of TCP/IP interactions. However, when link cost is a combination of both congestion price and a static component, analytic solution or even proof of the existence of an equilibrium is an open problem[12]. We thus focus on purely dynamic routing where the link cost is the congestion price. According to the KKT optimality condition

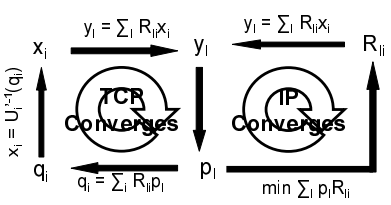


Fig. 1. Illustration of System Model One.

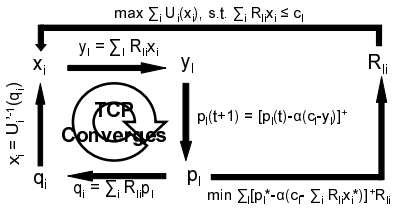


Fig. 2. Illustration of System Model Two.

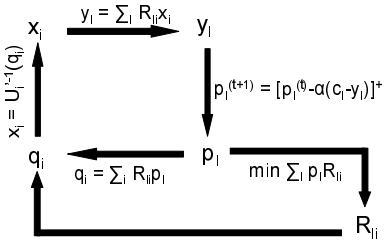


Fig. 3. Illustration of System Model Three.

(4), congestion price has to be zero when link load is strictly less than link capacity. Therefore, to avoid the case of random routing due to zero link costs, we need a topology and traffic model that can avoid zero congestion prices.

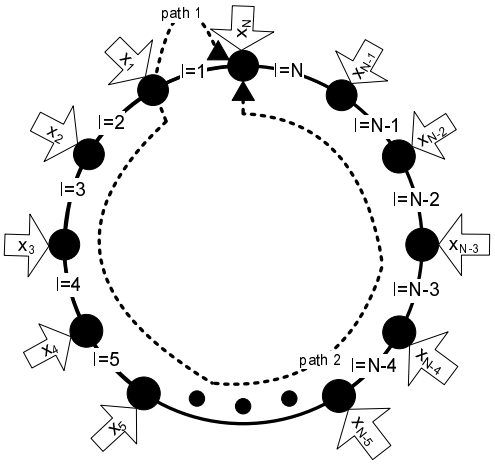


Fig. 4. N-node ring topology with N sources.

Consider a ring topology with N nodes, each of which being a source with a destination being the clockwise neighbor node as shown in Figure 4. Note that we can interchange l and i indices in this case. Each source has two possible paths: a one-hop path and an $(N-1)$-hop path. For the problem defined by (1) at optimality, the KKT conditions (4) allows for the constraint $Rx \le c$ being satisfied to be a potential solution. If R is invertible, then the constraint would be satisfied with equality and the source rates would be $x = R^{-1}c$. In addition, congestion prices would be non-zero and $p = qR^{-1}$, where $q_i = U_i'(x_i)$. There are degenerate cases where R is not invertible, *e.g.*, when two sources have the same split between paths. Those routing configurations would be changed in the next TCP/IP round since there would be at least one link with zero congestion price and the routing adaptation will change the routing matrix to take advantage of the zero-congestion-price link.

IV. Stability Analysis

In this section, stability analysis is performed on each System Model for the ring topology and traffic model described in Figure 4. We find that for System Model One, stringent initial conditions are required for convergence. For System Model Two, for small enough step size, convergence (to minimum-hop routing) is guaranteed. Recall that even for TCP to converge, α needs to be sufficiently small. For System Model Three, convergence (to minimum-hop routing) is guaranteed if there is a link whose capacity dominates those of other links, while other capacity configurations may also lead to converge (to non-minimum-hop routing).

A. Analysis of System Model One (Figure 1)

Each TCP/IP round consists of the following two loops:

- **TCP**: Complete iterations (2) and (3) to generate $x^*(t)$ and $p^*(t)$, where t indexes the iteration of the *joint* TCP/IP system.
- **IP**: Update the prices $p_l(k+1) = [p_l(k) - \alpha(c_l - y_l(k))]^+$, where k indexes the iteration *within* the IP loop. Then, for each source i, solve $\min_R \sum_l p_l(k)R_{li}$. The new R will update $y_l(k)$, which in turn updates $p_l(k+1)$.

We first present simple examples illustrating three possible system behaviors.

1. TCP/IP stable: Consider a three-node ring topology with unit capacity on all links, starting with shortest-path routing. Then the TCP/IP system converges to $x^* = [1\ 1\ 1]$; $p^* = [1\ 1\ 1]^T$; $R^* = R(0)$ and it is stable.

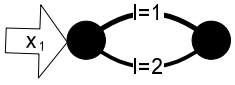


Fig. 5. Two-node topology with a single source.

2. IP unstable: From Figure 5, there are two parallel links with unit capacity and only one source-destination pair. Let $a = 0.5 + \epsilon$, $b = 1 - a$ and $c = b/a$, where $0 < \epsilon \ll 1$. Given $R(0) = [a\ b]^T$, TCP converges to $x^*(0) = 1/a$, $p^*(0) = [1\ 0]^T$. Inside the IP loop, each successive iteration produces: $p(k) = [1 - k\alpha, k\alpha c]^T$; $R(k) = [0\ 1]^T$, until $1 - k\alpha < k\alpha c$, at which point we have $R(k+1) = [1\ 0]^T$. This, however, triggers the congestion price of the top link to decrease with each iteration while the congestion price of the bottom link to increase until routing $R = [0\ 1]^T$. So in this case, the IP loop itself never converges.

3. IP stable, TCP/IP unstable: This example uses the same topology as example two. Initially, the top path is chosen, *i.e.*, $R(0) = [1\ 0]^T$. From TCP, $x^*(0) = [1]$; $p^*(0) = [0\ 0]^T$. The IP iteration converges to $R^* = [0\ 1]$; $p^* = [1\ 0]^T$ since all the traffic will be routed to the path with the lower

congestion price. In the next TCP iteration, however, $x^*(1) = [1]$; $p^*(1) = [0\ 1]^T$. It is easy to see the system ends up oscillating between routing on the top path and routing on the bottom path and never converges.

Theorem 1: For the ring topology and traffic model in Figure 4, System Model One converges (and necessarily to minimum-hop routing) if and only if the initial routing is minimum-hop routing on at least $N-1$ nodes.

Proof: Inside the IP loop, $p(k+1) = p(k)$ if all links are fully utilized. It is also easy to see that shortest-path routing in the IP loop means that each source does a comparison between its two paths, with three possibilities:

1) If $p_l < \sum_{j \neq l} p_j$, then choose the one-hop path.
2) If $p_l = \sum_{j \neq l} p_j$, then split arbitrarily between the two paths since the problem has many optimizers.
3) If $p_l > \sum_{j \neq l} p_j$, then choose the longer-hop path.

Since $p_l(1) > 0,\ \forall l$, for any source, if $p_l \geq \sum_{j \neq l} p_j$ for some link l, then all other sources must be doing minimum-hop routing. So there are only three possible routing configurations:

1) All sources choosing one-hop paths.
2) $N-1$ sources choosing one-hop paths, one splitting.
3) $N-1$ sources choosing one-hop paths, one source going on the longer-hop path. This is an unstable configuration.

Let $\mathcal{R}^{\mathcal{IP}}$ be the set of all routing configurations IP can produce. For the "if direction" of the theorem: If $R(0) \in \mathcal{R}^{\mathcal{IP}}$, then TCP will generate a source rate which fully utilizes all links under such a routing configuration. Inside the IP loop, shortest-path routing would produce $R = R(0)$, and the TCP/IP system is stable. For the "only if direction": If $R(0) \notin \mathcal{R}^{\mathcal{IP}}$, then TCP will generate a source rate which cannot fully utilize all links for $R \in \mathcal{R}^{\mathcal{IP}}$. Then inside the IP loop, there will be always be links with zero congestion price, and the IP loop will never converge. ■

Note that the stable solution is not necessarily optimal. As a simple example, consider $N = 3$, where link 1 has capacity 0.1 while all other links have unit capacity. Utilities are log functions. Minimum-hop routing achieves an aggregate utility of $\log 0.1$. If x_1 is split to have $2/11$ on the one-hop path and $9/11$ on the longer-hop path, however, then a higher aggregate utility of $3 \log 0.55$ can be achieved.

B. Analysis of System Model Two (Figure 2)

Due to special properties of the ring topology and traffic model, as explained in Section III.C, when routing only iterate once, after a few system iterations, it is safe to assume the congestion price is nonzero on every link. We can choose $\alpha < \max_l p_l^*/c_l$ to ensure $p_l' > 0,\ \forall l$, for all subsequent iterations. The optimization problem thus becomes:

$$\begin{array}{ll} \text{minimize} & \sum_l \left[p_l^*(t) - \alpha \left(c_l - \sum_k x_k^* \sum_{j'} H_{lj'}^i w_{j'}^i \right) \right] \\ & \times \sum_j H_{lj}^i w_j^i,\ \forall i \\ \text{subject to} & w_j^i \geq 0,\ \forall i, j;\ \sum_j w_j^i = 1,\ \forall i. \end{array} \tag{7}$$

Theorem 2: For the ring topology and traffic model in Figure 4, TCP/IP System Model Two converges (and necessarily to minimum-hop routing) if the step size is sufficiently small.

Proof: We can rewrite (7) as follows:

$$\begin{array}{ll} \text{minimize} & \alpha w^T X H^T H w + s^T w \\ \text{subject to} & A^T w = b \\ & w \succeq 0 \end{array} \tag{8}$$

where the symbols are defined below. H is simply the topology matrix. Construct a stacked-up version of w as $w = [w_1^1\ w_2^1\ w_1^2\ w_2^2\ w_1^3\ w_2^3\ \cdots\ w_1^N\ w_2^N]^T$. X is a $2N \times 2N$ matrix where row $2i$ and row $2i+1$ are filled with x_{i+1} for $i = 0$ to $2N-1$, *i.e.*,

$$X = \begin{bmatrix} x_1 & x_1 & x_1 & x_1 & \cdots & x_1 \\ x_1 & x_1 & x_1 & x_1 & \cdots & x_1 \\ x_2 & x_2 & x_2 & x_2 & \cdots & x_2 \\ x_2 & x_2 & x_2 & x_2 & \cdots & x_2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_N & x_N & x_N & x_N & \cdots & x_N \\ x_N & x_N & x_N & x_N & \cdots & x_N \end{bmatrix},$$

s is the linear term of the optimization objective and it depends on p_l and c_l:

$$-s = \begin{bmatrix} \alpha c_1 - p_1^* \\ (N-1)(\alpha c_1 - p_1^*) \\ \alpha c_2 - p_2^* \\ (N-1)(\alpha c_2 - p_2^*) \\ \vdots \\ \alpha c_N - p_N^* \\ (N-1)(\alpha c_N - p_N^*) \end{bmatrix}.$$

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}^T,$$

$$b = [1\ 1\ \cdots\ 1]^T.$$

This is an equality-constrained convex quadratic minimization problem where the KKT optimality conditions can be written as a system of linear equations:

$$\begin{bmatrix} \alpha X H^T H & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} w^* \\ v^* \end{bmatrix} = \begin{bmatrix} -s \\ b \end{bmatrix} \tag{9}$$

Solving for w_1^{i*} through matrix inversion, we obtain

$$w_1^i = \frac{N-1}{N-2} - a_N \left(\frac{c_i}{x_i} - \frac{p_i}{\alpha x_i} \right) - b_N \sum_{j \neq i} \left(\frac{c_j}{x_j} - \frac{p_j}{\alpha x_j} \right)$$

where $a_N = \frac{1}{4} + \frac{4-N}{4(N-2)^2}$, $b_N = \frac{4-N}{4(N-2)^2}$.

Projecting the above solution to the nonnegative quadrant, we map all $w_1^i > 1$ to 1 and all $w_1^i < 0$ to 0.

Lemma 1: $w_1^{i*} = 1,\ \forall i$, is a stable solution.

Proof: Since TCP will produce $x_1 = c_1, x_2 = c_2, x_3 = c_3, \ldots, x_N = c_N$, and all the links will be fully utilized. Then the routing adaptation will result in $w_1^i = 1 + x p_i/(\alpha x_i) + y \sum_{j \neq i} p_j/(\alpha x_j)$, which implies $w_1^{i*} = 1,\ \forall i$. ■

For convergence to minimum-hop routing, the following must hold:

$$\frac{1}{N-2} \geq a_N \left(\frac{c_i}{x_i} - \frac{p_i}{\alpha x_i} \right) - b_N \sum_{j \neq i} \left(\frac{c_j}{x_j} - \frac{p_j}{\alpha x_j} \right) \quad (10)$$

There are two cases depending on the size of the ring:

1) $N \leq 4$: $a_N \geq 0, b_N \geq 0$ and for sufficiently small α, (10) will hold.
2) $N > 4$: $a_N \geq 0, b_N < 0$, in this case the b_N term helps with achieving inequality (10). A sufficiently small α (but bigger than the largest α allowed in the $N \leq 4$ case) will enable (10) to hold.

■

As in the previous section, we note that minimum-hop routing is not necessarily the optimal solution.

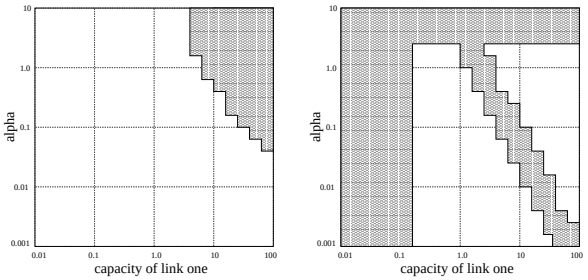
C. Analysis of System Model Three (Figure 3)

Theorem 3: For the ring topology and traffic model in Figure 4, TCP/IP System Model Three converges to minimum-hop routing if the capacity of one link in the ring is sufficiently large and the step size is sufficiently small.

Proof: System Model Three can only converge to (x^*, R^*) if the congestion prices after a certain time index evolve to maintain R^*. If the R is constant, System Model Three reduces to a TCP loop, and will converge to the optimal x^* for the given R. Without a loss of generality, we may assume after a number of iterations, at least one link becomes congested, then, following directly from the analysis of System Model One, there is at most one source splitting or going on the longer-hop path. Let the potentially splitting source be node 1 and let $a = w_1^1$, $0 \leq a \leq 1$, parameterize all possible R. It follows $y_1(t) = \frac{a}{ap_1(t)+(1-a)\sum_2^N p_l(t)}$; $y_l(t) = \frac{1-a}{ap_1(t)+(1-a)\sum_2^N p_l(t)} + \frac{1}{p_l(t)}, l \neq 1$. There are three cases:

1) One source taking longer-hop path ($a = 0$): $p_1(t+1) = [p_1(t) - \alpha c_1]^+$, this is a monotonically decreasing function, so after a number of iterations, $p_1 > \sum_2^N p_l$ will no longer hold and R will change.
2) One source splitting ($0 < a < 1$): Convergence requires $p_1(t+k) = \sum_2^N p_l(t+k), \forall k > 0$, given $p_1(t) = \sum_2^N p_l(t)$. This holds when $(c_1 - \sum_2^N c_l) = (y_l - \sum_2^N y_l)$. Since y_l is changing with change in R and x, while c stay constant. This configuration is unstable.
3) All sources taking one-hop path ($a = 1$): Convergence requires $p_1(t+1) < \sum_2^N p_l(t+1)$ given $p_1(t) < \sum_2^N p_l(t)$. Since $\sum_2^N p_l(t+1) - p_1(t+1) = (\sum_2^N p_l(t) - p_1(t))(1 - \alpha(\sum_2^N p_l^{-2}(t) - p_1^{-2}(t))) + \alpha(c_1 - \sum_2^N c_l)$. If α is chosen sufficiently small so that $1 > \alpha(\sum_2^N p_l^{-2}(t) - p_1^{-2}(t))$ and the capacity distribution is such that $c_1 > \sum_2^N c_l$, then $p_1(t+1) < \sum_2^N p_l(t+1)$ is guaranteed.

In summary, for sufficiently small step size and a capacity distribution dominated by c_1, convergence to minimum-hop routing is guaranteed. ■



(a) System Model Two (b) System Model Three

Fig. 6. Convergence (white) and divergence (shaded) for five-node ring

V. Simulation Results

First, we simulate over the ring topology for all three systems to confirm our stability analysis results. Secondly, the achieved aggregate utility is compared to the TCP/IP joint optimum. The results demonstrate that increased homogeneity and faster timescale interactions shrink the gap to joint optimum. Finally, an access-core network topology is simulated for System Models Two and Three. We use log-utility in all our simulations. In all plots, the x-axis is capacity of link 1 of the ring, shown on a log scale.

We use a combination of Matlab and MOSEK (www.mosek.com) environments to numerically study the interactions of TCP congestion control and IP routing. Most of the implementation is straight-forward, except for the joint optimization problem (5) that is a non-convex optimization in (x, R). With a simple change of variable $y^i = x_i w^i$, however, the problem can be transformed to a convex optimization problem in y:

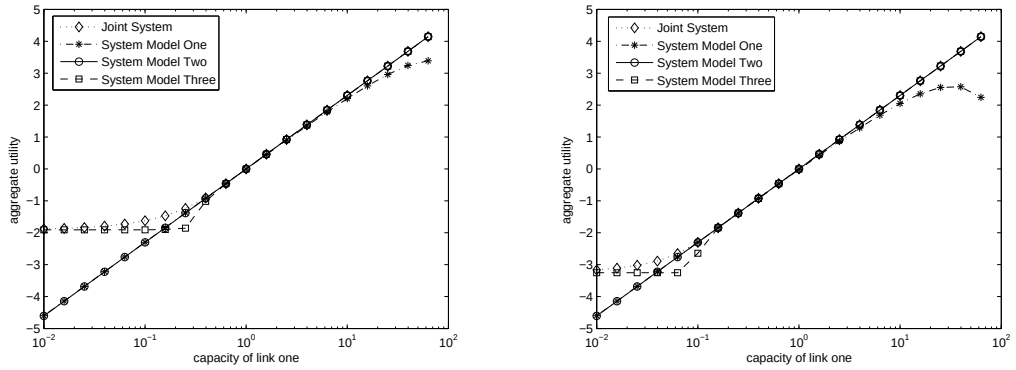
$$\begin{array}{ll} \text{maximize} & \sum_i U_i(\mathbf{1}^\mathbf{T} y^i) \\ \text{subject to} & Hy \leq c \\ & y \geq 0. \end{array}$$

A. Stability of Ring Topology

Only System Model Two and Three are shown, because System Model One's convergence depends heavily on the initial routing configuration. In Figure 6, the shaded region represents the divergent region. Figure 6(a) confirms our findings that a smaller step size helps with convergence for System Model Two. Figure 6(b) also confirms the analytic results in showing that, for a certain capacity range, convergence is very difficult, for the other capacity range, smaller step size helps with convergence. For the ring topology and traffic model in Figure 4, stability is certainly dependent on timescale as the attraction regions for the System Models are quite different. System Model Two appears to be the best timescale interaction for a stable solution. Having a more dynamic interaction (System Model Three) or a more static interaction (System Model One) reduces stability.

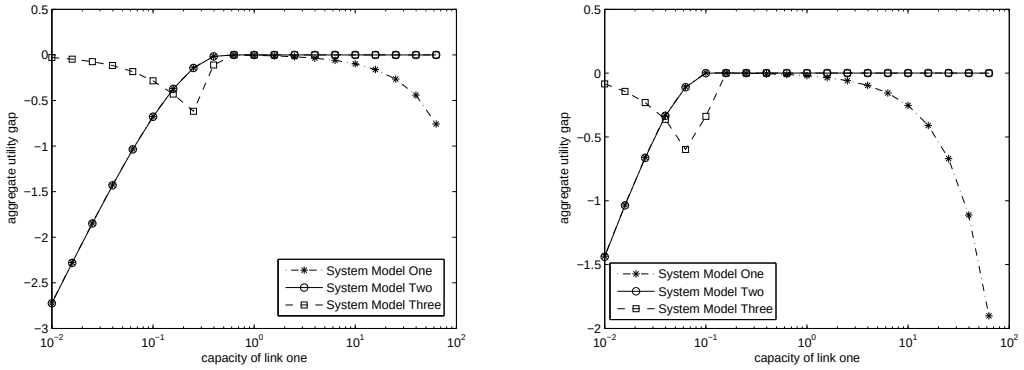
B. Optimality of Ring Topology

In this section, we examine the optimality gap of each System Model (at a stable point). For System Model One, we



(a) Three-node ring ($N = 3$) (b) Ten-node ring ($N = 10$)

Fig. 7. Aggregate utility for optimal solution and the three System Models



(a) Three-node ring ($N = 3$) (b) Ten-node ring ($N = 10$)

Fig. 8. Aggregate utility *gap* for the three System Models

assume the initial routing is such that source 1 is split 99.5% on the one-hop path and 0.5% on the $(N-1)$-hop path. In the plots in Figure 7, the dotted line signifies the joint optimum (solution to the joint optimization problem (5)), and the other lines represent the three System Models. It can be seen that while capacity of link one is close to that of the other links, *i.e.*, the system is homogeneous, there is no utility gap between the distributed and joint system. This holds for both the three-node ring and the ten-node ring cases. As to be expected, the effect of heterogeneity is higher for the three-node ring since the standard deviation for the distribution would be higher for the same value of capacity on link one. The effect of link capacity homogeneity is best seen in Figure 8, where the difference between each system and the joint optimum is plotted. All four figures demonstrate that the more dynamic the timescale interaction, the closer a System Model achieves the joint optimum when it converges to a stable point.

C. Stability and Optimality of Access-Core Topology

We next simulate over a tree-mesh topology, *e.g.*, in Figure 9, to gain further insights on behaviors of joint system models for access-core type of topology. In the middle is a full mesh representing the core of the network with rich connectivity. On the edge are three access tree subnetworks. There are six possible source nodes and twelve possible source-destination pairs. Of the twelve pairs, $1-3, 1-5, 2-4, 2-6, 3-5, 4-6$ are chosen, and for each source-destination pair, the three minimum-hop paths are chosen as possible paths. The simulations were performed by assuming the capacity of the links follows a truncated (so as to avoid negative values) Gaussian distribution, with an average of 100 and a standard deviation that we vary from 0 to 50. Ten realizations at each standard deviation are tested. System Model One is not simulated since it does not converge except under stringent initial conditions.

System Model Two converges for the range of step size from 0.01 to 100. It has a significant gap from optimality, however, as can be seen in Figure 10 where each individual experiment is shown with an x and the solid line indicates the averages. From the solid line, it is easy to observe that, once again, 'homogeneity helps attaining optimality'. For System Model Three, the simulations (graphs not shown) show that it is prone to being stuck in an infeasible region for a large range of step sizes. In such cases, at each routing update, routing swings from one configuration to another, which in turn causes the link utilization to swing from one infeasible point to another,

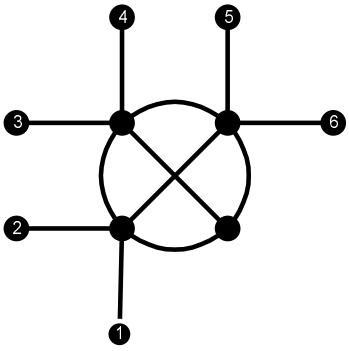


Fig. 9. An access-core network topology.

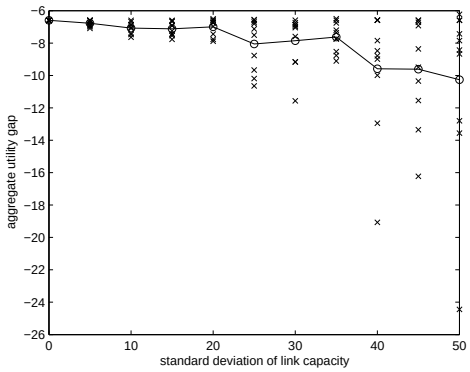


Fig. 10. Aggregate utility gap for access-core network, System Model Two.

causing constant congestion, route oscillations and packet loss.

VI. Conclusions and Future Work

While congestion price is used by TCP for distributed congestion control and may seem to be a natural choice of link weights for dynamic routing, it is prone to oscillations if deployed in practice. In particular, for stability in a ring topology, stringent initial conditions are required for System Model One and specific capacity configurations are required for System Model Three. Even when the joint system does converge, there exists large optimality gap for realistic topologies. Using terminology in the unifying framework of "Layering As Optimization Decomposition" [15], congestion price is a poor "layering price" for TCP/IP interaction. Compared to all other cross-layer designs based on "Layering As Optimization Decomposition", this is so far the only exception where congestion price (or queuing delay) is not an appropriate coordination across layers. While we have not addressed stochastic traffic or feedback delay issues [16], [17] in our model, it is unlikely that such features in the model would enhance stability of the TCP/IP system.

There are several directions for future work. To avoid instability of TCP/IP joint system, we can either adopt the heuristics of adding a static component to the link weight (as in the early ARPANET work [9]), or change the feedback metric and route optimization problem. For example, in current traffic engineering practice, routing would be trying to *centrally* minimize a *penalty function* of link utilization based on a network-wide view of the current offered traffic [7]. Turning from analysis to design, we can also define an optimization where a weighted difference of end-user utilities and network operator penalty function is maximized over both routes and source rates that are constrained by link capacities. A distributed solution to this problem and its implementation over existing TCP and traffic engineering systems have recently been presented [18].

Acknowledgment

We would like to thank Steven Low, Jiantao Wang, Lun Li and Ao Tang of Caltech for illuminating discussions on this topic. This work has been supported in part by NSF grants CNS-0519880 and CCF-0448012, and a Cisco University Research Program grant.

References

[1] F. P. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: Shadow prices, proportional fairness and stability," *J. of Operational Research Society*, vol. 49, pp. 237–252, March 1998.

[2] R. J. La and V. Anantharam, "Utility-based rate control in the internet for elastic traffic," *IEEE/ACM Trans. Networking*, vol. 10, pp. 272–286, April 2002.

[3] S. H. Low, L. Peterson, and L. Wang, "Understanding Vegas: A duality model," *J. of the ACM*, vol. 49, pp. 207–235, March 2002.

[4] S. H. Low, "A duality model of TCP and queue management algorithms," *IEEE/ACM Trans. Networking*, vol. 11, pp. 525–536, August 2003.

[5] R. Srikant, *The Mathematics of Internet Congestion Control*. Birkhauser, 2004.

[6] B. Fortz and M. Thorup, "Optimizing OSPF weights in a changing world," *IEEE JSAC*, vol. 20, pp. 756–767, May 2002.

[7] J. Rexford, "Route optimization in IP networks," in *Handbook of Optimization in Telecommunications*, Springer Science + Business Media, February 2006.

[8] D. Bertsekas, "Dynamic behavior of shortest-path routing algorithms for communication networks," *IEEE Trans. Automatic Control*, pp. 60–74, February 1982.

[9] J. M. McQuillan and D. C. Walden, "The ARPA network design decision," *Computer Networks*, vol. 1, pp. 243–289, August 1977.

[10] Z. Wang and J. Crowcroft, "Analysis of shortest-path routing algorithm in dynamic network environment," *ACM SIGCOMM Computer Communication Review*, vol. 22, pp. 63–71, April 1992.

[11] E. J. Anderson and T. E. Anderson, "On the stability of adaptive routing in the presence of congestion control," in *Proc. IEEE INFOCOM*, April 2003.

[12] J. Wang, L. Li, S. H. Low, and J. C. Doyle, "Cross-layer optimization in TCP/IP networks," *IEEE/ACM Trans. Networking*, vol. 13, pp. 582–595, June 2005.

[13] J. Moy, "OSPF Version 2." RFC 2328, April 1998.

[14] E. Rosen, A. Viswanathan, and R. Callon, "Multiprotocol Label Switching Architecture." RFC 3031, January 2001.

[15] M. Chiang, S. H. Low, R. A. Calderbank, and J. C. Doyle, "Layering as optimization decomposition." To appear in *Proceedings of IEEE*, 2006. Shorter version appeared in *Proc. Conf. Inform. Science and Sys.*, March 2006.

[16] X. Lin and N. B. Shroff, "Utility Maximization for Communication Networks with Multi-path Routing," *IEEE Trans. Automatic Control*, 2006. To appear.

[17] F. Kelly and T. Voice, "Stability of end-to-end algorithms for joint routing and rate control," *ACM SIGCOMM Computer Communication Review*, vol. 35, pp. 5–12, April 2005.

[18] J. He, M. Chiang, and J. Rexford, "DATE: Distributed Adaptive Traffic Engineering." Poster session at INFOCOM 2005.