On the Resource Utilization and Traffic Distribution of Multipath Transmission Control

UMass Computer Science Technical Report UM-CS-2011-005

Bo Jiang[1], Yan Cai[2], Don Towsley[1]

[1] {bjiang, towsley}@cs.umass.edu [2] ycai@ecs.umass.edu

University of Massachusetts, Amherst, MA, USA

Abstract

There is growing interest in the development and deployment of multipath rate and route control mechanisms for the Internet, due to their ability to exploit bandwidth resources, alleviate network congestion, and provide robustness against failures. However, two performance issues have been uncovered: low link utilization when the number of flows is small, and route flappiness, namely the traffic of a flow tends to concentrate on one path and then another. In this paper we study these issues with respect to several variations of multipath rate and route control algorithms. We demonstrate the qualitatively different impacts that the couplings of the increase and decrease phases have on link utilization. We also demonstrate how the coupling strength affects both the long-term and short-term traffic distributions among different paths. In particular, we show that the flappy behavior is prominent only when there is strong coupling in both the increase and decrease phases, and when the number of good paths is small.

1 Introduction

There is growing interest in the development and deployment of multipath rate and route control mechanisms for the Internet. Such mechanisms allow flows to transfer large data files over multiple paths and control the data rates over these paths in response to network congestion and failures [17, 6, 5, 11, 14, 18, 19]. This has a number of advantages over single path rate control as found in TCP. First, they provide robustness against failures or onset of congestion; flows can shift data from the failed or congested paths to the remaining paths. Second, in a network where most/all flows use such mechanisms, network capacity can substantially increase and the task of traffic engineering substantially decrease [10].

A growing body of work [6, 5, 9, 17, 13, 16, 19] focuses on the design of such mechanisms. One approach

simply relies on setting up and maintaining a TCP connection over each path and feeding data into each of these at the maximum rate that it can support. Examples of this approach include http transfer applications, such as DownThemAll [4], and peer-to-peer applications, such as BitTorrent [3]. The former uses a number of TCP connections to transfer a large data file simultaneously and the latter maintains a number of, typically four, active connections to other peers with an additional path periodically chosen at random together with a mechanism that retains the best paths (as measured by throughput).

However, such an approach has been shown not to achieve the full potential of multipath [10]. Furthermore, it disregards single path TCP connections and treats them unfairly. Thus the second approach [6, 5] designs controllers that coordinate the sending rates over each path. These controllers can achieve high aggregate throughput while at the same time being fair to single path flows. We shall refer to this class of multipath rate and route control as multipath transmission control (MTC).

Several proposals for multipath rate and route controllers have been made recently [17, 5, 6, 9, 18, 19]. These proposals find their root in the resource-sharing theoretic framework established by Kelly et al. [7]. They have shown that distributed congestion control mechanisms such as those found in TCP maximize the sum of flow utilities, where the utility function defines a fairness criterion among the flows. Moreover, the framework accommodates not only single path flows but also multipath flows.

The inherent feature of multipath routing also benefits security by making it more difficult to eavesdrop the entire transferred data [12, 13]. [12] proposes a distributed secure multipath approach to protect data transfers on the Internet by dispersing data across multiple paths. [13] proposes a secure data delivery mechanism, named SPREAD, to enhance the security of data transfers in mobile ad hoc networks.

Most of the work mentioned above has been theoretical in nature. Recently, attempts have been made to put these ideas into practice. Unfortunately these have uncovered two problems: low resource utilization and flappiness. [11] studied the performance of an MTC algorithm based on proportional fairness, which achieves lower resource utilization than predicted by fluid models used in theoretical studies [6, 5] in the case of a small number of flows. Second, experimental studies [18, 16] show that a multipath flow exhibits *flappy* behavior, namely the traffic tends to shift from one subflow to another from time to time. As a result, the flow transfers data over only one of its paths for most of the time. This kind of behavior reduces the robustness of the mechanism. Failures or the onset of congestion affect a flow much more when the data is being transferred over only one path rather than spread over all the paths available to the flow.

Our work is motivated by these performance issues, which necessitate a careful study of the coupling mechanisms of MTC algorithms, namely, how they coordinate the sending rates over different subflows. In this paper we study the multiplicative-increase and multiplicative-decrease (MIMD) MTC algorithm investigated in [11] as well as a family of additive-increase and multiplicative-decrease (AIMD) MTC algorithms.

Our contributions in this paper are three-fold:

1. We provide a quantitative explanation for the low link utilization observed in previous studies. Our analysis also reveals the qualitatively different natures of the couplings in the increase and decrease phases of MTC algorithms. Coupling in the increase phase reduces the link utilization by changing the window trajectory from linear to convex, while coupling in the decrease phase simply incurs a larger decrement of the window size upon a loss event.

2. We show how the coupling strength affects the long-term traffic distribution among paths of different loss probabilities. The stronger the coupling, the more traffic is concentrated on less congested paths.

3. We analyze the impacts of different parameters on the flappiness of an MTC flow. We show that the flappy behavior is prominent only when there is strong coupling in both the increase and decrease phases, and when the number of good paths is small.

The rest of the paper is organized as follows. Section 2 introduces the MTC controllers studied in this paper. Section 3 studies the link utilization of MTC flows. Section 4 analyzes the traffic distributions of MTC flows. Section 5 extends our studies using simulation. Related work is introduced in Section 6, and Section 7 concludes the paper.

2 Multipath Transmission Control

In this section we introduce the multipath transmission control (MTC) algorithms studied in this paper.

In general, an MTC flow refers to a network connection that transfers data between a sender and receiver pair using n subflows, as shown in Figure 1. The n subflows use n' ($\leq n$) distinct physical paths, which may partially overlap. The use of multiple paths is referred to as path diversity: the larger the number of underlying paths, the greater the path diversity. Path diversity is a hallmark of MTC that represents the potential of an MTC flow to explore network bandwidth. The subflows of an MTC flow are coupled to ensure fairness [2] between MTC flows and single path flows. Our goal is to study the coupling mechanisms in this paper.

While previous research has focused on rate-based MTC controllers [5, 6], the controllers we study in this paper are window-based. For an MTC flow consisting of n subflows, let $w_i(t)$ denote the congestion window size of the i-th subflow at time t, and $w(t) = \sum_{i=1}^{n} w_i(t)$ the aggregate congestion window size.

The MTC controllers we study in this paper can be classified as either MIMD or AIMD based, according to how they react to the absence and presence of losses, which are signaled using ACKs and NAKs, respectively.

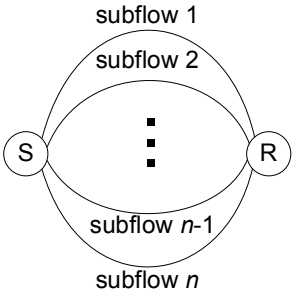
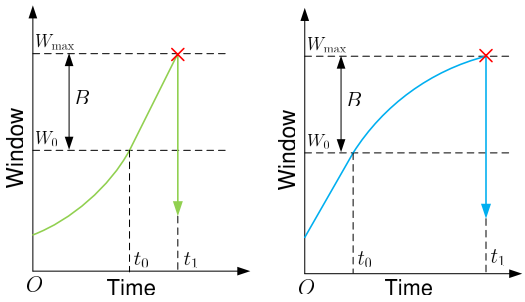


Figure 1: An MTC flow consisting of n subflows.



(a) STCP (b) TCP

Figure 2: Typical cycles of (S)TCP congestion window trajectory. Packets are dropped when the window exceeds the maximum supportable size $W_{\max}$.

MSTCP is MIMD based and corresponds to proportional fairness. It is the multipath version of Scalable TCP [8]. The window update rule is as follows,

$$w_i^+ = \begin{cases} w_i^- + a, & \text{if ACK,} \\ \max\left\{w_i^- - bw^-, 1\right\}, & \text{if NAK,} \end{cases} \tag{1}$$

where w_i^- and w_i^+ represent the window size of the i-th subflow immediately before and after the update, respectively, and $w^- = \sum_{i=1}^n w_i^-$. The congestion window $w_i(t)$ of the i-th subflow is increased by a constant a upon receipt of an ACK and decreased by $\min\{bw^-, w_i^- - 1\}$ upon receipt of a NAK. This MTC variation was first investigated in [11], where the parameters $a = 0.03$ and $b = 1/2$ were used.

The other MTC controllers are AIMD based and their window update rules have the following general form,

$$w_i^+ = \begin{cases} w_i^- + \frac{a}{(1-\alpha)w_i^- + \alpha w^-}, & \text{if ACK,} \\ \max\left\{w_i^- - b[(1-\beta)w_i^- + \beta w^-], 1\right\}, & \text{if NAK.} \end{cases} \tag{2}$$

The parameters α, β $(0 \le \alpha, \beta \le 1)$ control the strength of couplings among the subflows in the increase and decrease phases, respectively, with larger values corresponding to stronger couplings.

This class of controllers contains the multipath versions of TCP and will be referred to as MTCP. When $n = 1$, $a = 1$, and $b = 1/2$, they all reduce to TCP Reno [1]. The following three variations of MTCP are of particular interest.

COUPLED DECREASE MTCP corresponds to $\alpha = 0$ and $\beta = 1$. The congestion window $w_i(t)$ is increased by a/w_i upon receipt of an ACK and decreased by $\min\{bw, w_i - 1\}$ upon receipt of a NAK.

COUPLED INCREASE MTCP corresponds to $\alpha = 1$ and $\beta = 0$. The congestion window $w_i(t)$ is increased by a/w upon receipt of an ACK and decreased by $\min\{bw_i, w_i - 1\}$ upon receipt of a NAK.

FULLY COUPLED MTCP corresponds to $\alpha = 1$ and $\beta = 1$. The congestion window $w_i(t)$ is increased by a/w upon receipt of an ACK and decreased by $\min\{bw, w_i - 1\}$ upon receipt of a NAK.

3 Link Utilization

It was observed in [11] that when the number of simultaneous flows is small, the controller in (1) leads to low link utilizations due to the back-off being proportional to the total congestion window upon receipt of a NAK. In this section, we first provide a more quantitative explanation of this phenomenon through a deterministic analysis of the extremal case where there is only one MSTCP flow. We then extend our analysis to MTCP and multiple simultaneous flows.

Throughout this section, we will assume full path diversity for an MTC flow, i.e. different subflows use different paths that do not share the same bottleneck link. We will also assume all paths are homogeneous. These assumptions allow us to study the key characteristics of the coupling mechanisms without overly complicating the analysis.

Let n denote the number of subflows in an MTC flow. Denote by D, B, and C, respectively, the common values of the 2-way propagation delay, bottleneck buffer size, and bottleneck link capacity of the n paths.

The link utilization U_i of the i-th path is the ratio of the throughput T_i to the capacity C of the i-th path, i.e. $U_i = T_i/C$. The overall link utilization U is the ratio of the total throughput $T_{\text{tot}} = \sum_{i=1}^{n} T_i$ to the total capacity $C_{\text{tot}} = nC$, so $U = n^{-1} \sum_{i=1}^{n} U_i$, which is the average of the link utilizations of each path.

The round-trip time (RTT) of the i-th path at time t is

$$\tau_i(t) = \max\left\{\frac{w_i(t)}{C}, D\right\}, \tag{3}$$

and the data rate is

$$r_i(t) = \min\left\{\frac{w_i(t)}{D}, C\right\} = \frac{w_i(t)}{\tau_i(t)}. \tag{4}$$

Note that the minimum window size W_0 for the data rate to reach full capacity is given by the capacity-delay product, $W_0 = CD$. The maximum supportable window size $W_{\max}$ of a path is $W_{\max} = W_0 + B = (1 + \gamma)W_0 = \kappa W_0$, where $\gamma = B/W_0$ and $\kappa = 1 + \gamma$.

Throughout the rest of this section, we make the following assumption.

Assumption 1 (Deterministic drop)**.** *A packet loss on the i-th path occurs if and only if its congestion window w_i reaches the maximum supportable size $W_{\max}$.*

Figure 2 shows typical cycles of the trajectory of TCP/STCP congestion window. A packet loss occurs when the window reaches the maximum supportable size, and then the window size is decreased accordingly.

During each cycle, i.e. between two consecutive drops, we model the window as fluid, which is commonly used, and describe its dynamics using ordinary differential equations (ODEs).

3.1 MSTCP

In this subsection, we analyze the link utilization of the MSTCP controller described by (1). We first focus on a single path in a single cycle between two consecutive losses. By doing so, we have decoupled the behavior of the increase phase of the congestion control mechanism from the interflow interactions. We will then analyze the interactions among subflows under additional assumptions.

Consider a cycle on the i-th path. The dynamics of the window $w_i(t)$ can be described by the ODE,

$$\frac{d}{dt}w_i = a\frac{w_i}{\tau_i} = \begin{cases} \frac{a}{D}w_i, & \text{if } w_i \le W_0, \\ aC, & \text{if } w_i \ge W_0, \end{cases} \tag{5}$$

where (3) has been used. This behavior is illustrated in Figure 2(a).

Suppose $w_i(0) = W_{\max}$ and let the window size $w_i(0^+)$ immediately after the decrease be a fraction f of $w_i(0)$, i.e. $w_i(0^+) = fw_i(0)$. If $w_i(0^+) \le W_0$, then the window size will grow exponentially until it reaches W_0 at time t_0 (Figure 2(a)). It then grows linearly until it reaches $W_{\max}$ at time t_1, where

$$t_1 = \frac{W_0}{aC}\left(\frac{W_{\max}}{W_0} - 1 - \log\frac{W_0}{w_1(0^+)}\right).$$

By (4) and (5), the data rate is $r_i(t) = a^{-1}dw_i/dt$, so the throughput in the given cycle is

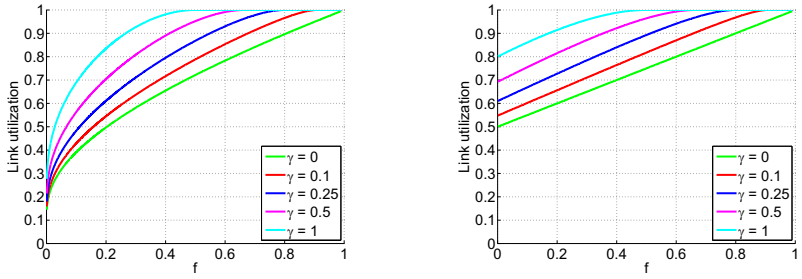
$$T_{\text{cycle}} = \frac{1}{t_1}\int_0^{t_1} r_i(t)dt = \frac{1}{t_1}\int_0^{t_1}\frac{1}{a}\frac{dw_i}{dt}dt = \frac{W_{\max} - w_i(0^+)}{at_1} = C\frac{\kappa(1-f)}{\kappa - 1 - \log(\kappa f)}.$$

If $w_i(0^+) \ge W_0$, then data is always sent at full capacity, so $T_{\text{cycle}} = C$. Therefore, the link utilization of a cycle with initial window size $fW_{\max}$ is given by

$$U_{\text{cycle}} = \frac{T_{\text{cycle}}}{C} = \begin{cases} \dfrac{\kappa(1-f)}{\kappa - 1 - \log(\kappa f)}, & \kappa f \le 1, \\ 1, & \kappa f \ge 1. \end{cases} \tag{6}$$

In general, different cycles have different values of f and different durations, despite the fact that the windows change deterministically over time under Assumption 1. The link utilization of the flow is then the average link utilization of all cycles of all paths weighted by the durations of the cycles.

Figure 3(a) plots the per-cycle link utilization (6) as a function of f. Note that the link utilization is very



(a) STCP (b) TCP

Figure 3: Per-cycle link utilization of STCP/TCP. Recall that $\gamma = B/W_0$ is the buffer size normalized by the capacity-delay product W_0.

low for small f. In fact, it goes to 0 as $f \to 0$. We will show next that the coupling in the decrease phase of the MSTCP controller in (1) leads to small values of f and hence low link utilizations, as observed in [11].

To get a sense of which values of f might be observed in practice, we make the following simplifying assumptions, which will be relaxed later in simulations.

Assumption 2. *The congestion windows of all subflows exhibit the same periodic pattern and are evenly spaced, that is, for any* $1 \le i \le n$,

$$w_i(t) = w_1\left(t - \frac{i-1}{n}L\right), \tag{7}$$

where L *is the period of* $w_i(t)$. *Note that* L/n *is the period of* $w(t)$. *By shifting the time axis, we assume the congestion window of the first path has drops at multiples of* L, *so* $w_1(t)$ *is continuous on* $(0, L]$ *but* $w_1(0) = W_{\max} > w_1(0^+)$. *Note that the value of* f *is the same for all cycles and is given by* $f = w_1(0^+)/W_{\max}$.

Assumption 3. *The buffer size is small relative to the capacity-delay product* W_0, *i.e.* $\gamma \approx 0$, *so the instantaneous RTT is a constant* $\tau_i = D$ *for* $i = 1, 2, \ldots, n$.

Under these assumptions, it is shown in Appendix A that

$$f = \max\{(1-b)^n, W_0^{-1}\}. \tag{8}$$

Since all cycles of all the paths are equivalent in this case, substituting (8) into (6) gives the overall link utilization U (see Figure 4). Note that f drops off exponentially as the number of subflows increases. For $b = 1/2$, it takes only $n = 3$ to bring f down to 0.125. Although assumption (7) almost never holds in practice due to its extreme sensitivity to timing, the point of this analysis is to show that, as the number of subflows increases, f drops off very fast and so does the link utilization.

3.2 MTCP

In this subsection, we analyze the link utilizations of the three versions of MTCP introduced at the end of Section 2. We will see that the couplings in the decrease and increase phases have qualitatively different impacts on link utilization.

3.2.1 Coupled Decrease

The analysis for the Coupled Decrease MTCP parallels that of Section 3.1. We first focus on a single path in a single cycle between two consecutive losses. During one cycle on the i-th path, the dynamics of the window $w_i(t)$ are described by the following ODE,

$$\frac{d}{dt}w_i = \frac{a}{\tau_i} = \begin{cases} a/D, & \text{if } w_i \le W_0, \\ aC/w_i, & \text{if } w_i \ge W_0, \end{cases} \tag{9}$$

where (3) has been used. This behavior is illustrated in Figure 2(b).

As shown in Appendix B, the link utilization of a cycle with initial window size $fW_{\max}$ is

$$U_{\text{cycle}} = \frac{T_{\text{cycle}}}{C} = \begin{cases} \dfrac{\kappa^2(1-f^2)}{\kappa^2+1-2\kappa f}, & \kappa f \le 1, \\ 1, & \kappa f \ge 1. \end{cases} \tag{10}$$

Figure 3(b) plots the per-cycle link utilization (10) as a function of f. The link utilization decreases with decreasing f. In contrast to Figure 3(a), however, it is lower bounded by 0.5. This is not surprising, since the data rate always increases linearly before it gets capped at the capacity, so its average can never fall below half of the capacity. We will no longer observe this if the increase phase is also coupled.

Under Assumptions 2 and 3, we establish (Appendix B)

$$f = \max\left\{\frac{2-b(n+1)}{2+b(n-1)}, \frac{1}{W_0}\right\}. \tag{11}$$

Note that when $n \ge n_0 = 2/b - 1$, we have $f = W_0^{-1}$, which is generally very small. For $b = 1/2$ as in conventional TCP, $n_0 = 3$. Comparing MSTCP to Coupled Decrease MTCP, we see that strong coupling in the decrease phase generally leads to low link utilizations. The impact on MSTCP is more detrimental due to the convexity of the window trajectory in the increase phase.

3.2.2 Coupled Increase

For Coupled Increase MTCP, if all the subflows always have the same window size, which is possible in principle, then the aggregate congestion window $w(t)$ will behave as if it were a single TCP, so the link utilization is given by (10) with $f = 1 - b$. In particular, the link utilization is approximately 75% with a small buffer. However, if the subflows are out of phase, the interactions become complicated when the buffer size is large. Under Assumptions 2 and 3, we show (Appendix C) that link utilization is given by

$$U = \frac{\kappa(1-f)(1+f^{1/n})}{2n(1-f^{1/n})}, \tag{12}$$

with $f = \max\{1-b, W_0^{-1}\}$. For very large n and $1 - b > W_0^{-1}$, (12) yields

$$U \approx \lim_{n\to\infty} \frac{\kappa b(1+(1-b)^{\frac{1}{n}})}{2n(1-(1-b)^{\frac{1}{n}})} = \frac{\kappa b}{-\log(1-b)} \tag{13}$$

If b is close to one, which corresponds to a large penalty upon a packet loss, (13) shows that the link utilization becomes extremely small as the number of subflows becomes large. However, for $b = 1/2$, corresponding to the more commonly used TCP Reno, Figure 4 shows that the link utilization of Coupled Increase MTCP does not decrease much from single path TCP. In fact, if b is not too large, (13) can be approximated by its first order Taylor polynomial $\kappa(1 - b/2)$, which is the link utilization of single path TCP. We expect the link utilization to be similar to that of single path TCP even when Assumptions 2 and 3 are violated, as is confirmed later by simulation. Thus coupling the subflows in the increase phase alone avoids the problem of low link utilization for reasonable values of b.

3.2.3 Fully Coupled

Section 3.2.2 has shown that coupling in the increase phase alone has limited impact on link utilization. On the other hand, Section 3.2.1 has shown that coupling in the decrease phase generally reduces the link utilization considerably, but it is always lower bounded by 50%. However, coupling in both phases can lead to extremely low link utilizations. Under Assumptions 2 and 3, the utilization is given by (12) with f given by (8). When the number of subflows is very large, the link utilization is

$$U \approx \frac{W_0 - 1}{W_0 \log W_0}.$$

Note that the link utilization decreases in W_0 for the entire range $W_0 \in (1, \infty)$.

Remark 1. Figure 4 compares the link utilizations of the three versions of MTCP with $b = 1/2$, under Assumptions 2 and 3. The Fully Coupled MTCP has the lowest link utilization of the three and can

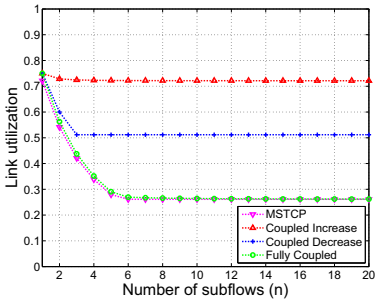


Figure 4: Link utilization versus number of subflows for MTCP under Assumptions 2 and 3; $b = 1/2$, capacity-delay product $W_0 = 42$.

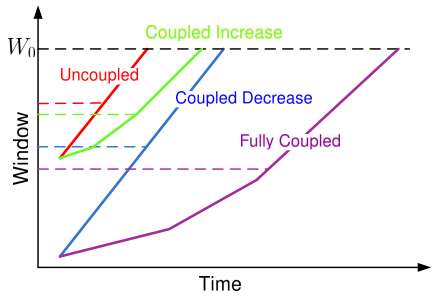


Figure 5: Trajectory of a typical cycle of the MTCP congestion window with buffer size zero. The dotted lines indicate the average heights.

have very small values as the number of subflows increases. Note also the close match of the curves for Fully Coupled MTCP and MSTCP. This is not surprising, since Appendix D.3 shows that the window trajectory for Fully Coupled MTCP is a broken line inscribed in an exponential curve, the hallmark of MSTCP.

Remark 2. Now we can summarize the impact of couplings on the link utilization of MTCP in a more geometric language. Figure 5 shows the trajectories of a typical cycle of the congestion window for an MTCP subflow. The link utilization is proportional to the average height of the solid curves. The coupling in the increase phase reduces the average by changing the trajectory from linear to convex. The coupling in the decrease phase reduces the average by dragging down the lower endpoint towards zero. When we have both, the impacts are amplified by each other. For MSTCP, the convexity is built-in without coupling in the increase phase, so coupling in the decrease phase alone has a major impact. When the buffer size is nonnegligible, the data rate will be capped at the capacity when the window size exceeds W_0, so the impacts of these couplings are weaker, but their qualitative natures are the same.

3.3 Multiple Flows

Recall that [11] observed low link utilization when the number of simultaneous flows is small. The previous two subsections have investigated this phenomenon in the extremal case of a single MTC flow. In this subsection, we consider multiple co-existing MTC flows, which presumably come from different users, and investigate how link utilization varies with the number of flows in a special case. We make Assumptions 2 and 3, and in addition, assume all the flows are identical in the sense that they use the same set of paths and are shifted versions of each other. More precisely, we have

Assumption 4. *Let m be the number of simultaneous flows and $w_i^j(t)$ be the congestion window on the i-th*

path of the j-th flow. Then for any $1 \le i \le n$, $1 \le j \le m$,

$$w_i^j(t) = w_i^1\left(t - \frac{j-1}{m}L\right), \tag{14}$$

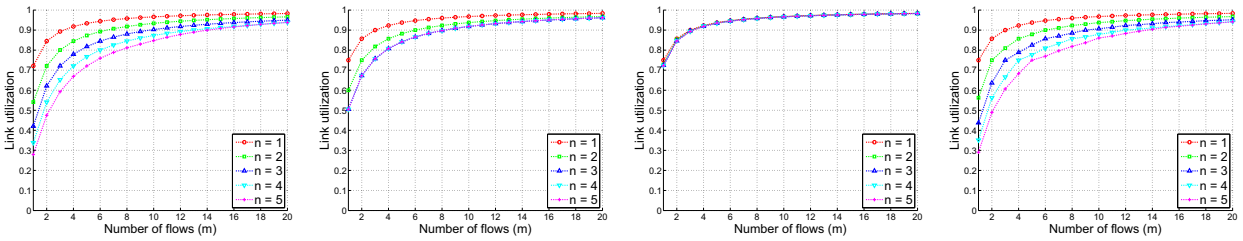
where L is the period of $w_i^j(t)$.

Note that Assumption 4 closely resembles Assumption 2. Thus much of the analysis for the single flow case carries over. The results are summarized in Table 1; see Appendix D for the derivation.

MSTCP	$U = \frac{m(1-f^{1/m})}{-\log f}$	$f = \max\{(1-b)^n, f^*\}$
Coupled Decrease MTCP	$U = \frac{m(1+f)}{m+1+(m-1)f}$	$f = \max\left\{\frac{2-b(n+1)}{2+b(n-1)}, \frac{m+1}{2W_0-m+1}\right\}$
Coupled Increase MTCP	$U = \frac{m(1-f)(1+f^{1/n})}{2n(1-f^{1/n})g(f,n,m)}$	$f = \max\{1-b, f^{**}\}$
Fully Coupled MTCP	$U = \frac{m(1-f)(1+f^{1/n})}{2n(1-f^{1/n})g(f,n,m)}$	$f = \max\{(1-b)^n, f^{**}\}$

Table 1: Link utilization for multiple flows. $f^* \in (0,1)$ is the root to $x^{1+1/m} - (1+W_0^{-1})x + W_0^{-1} = 0$; $g(f,n,m) = \sum_{j=1}^{m} f^{1-\frac{1}{n}\lceil\frac{jn}{m}\rceil}\left[1 - \left(\lceil\frac{jn}{m}\rceil - \frac{jn}{m}\right)\left(1-f^{1/n}\right)\right]$; $f^{**} \in (0,1)$ is the root to $g(f,n,m) - fW_0 = 0$.

Figure 6 shows the link utilization as a function of the number of simultaneous flows m. As m increases, we get better statistical multiplexing and hence larger link utilization. Note that Figure 6(a) closely resembles Figure 6(d). As observed in Remark 1, this is because the window trajectory for Fully Coupled MTCP is a broken line inscribed in an exponential curve, which is the window trajectory for MSTCP.



(a) MSTCP (b) Coupled Decrease MTCP (c) Coupled Increase MTCP (d) Fully Coupled MTCP

Figure 6: Link utilization for multiple identical MTC flows under Assumptions 3 and 4; $b = 1/2$, capacity-delay product $W_0 = 100$.

4 Traffic Distribution Among Subflows

In this section, we shift our attention to the impact of different couplings on the traffic distribution among different subflows of an MTC flow. Section 4.1 studies the long-term distribution through an equilibrium analysis. Section 4.2 examines the short-term distribution in terms of the so-called flappy behavior. We only consider the MTCP controller in (2) but allow arbitrary values of α and β and heterogeneous paths.

4.1 Equilibrium Analysis

In the presence of good statistical multiplexing on each path, the window dynamics of the MTCP controller in (2) are described by the following system of ODEs,

$$\frac{d}{dt}w_i = \frac{w_i}{\tau_i}\left(\frac{a(1-p_i)}{(1-\alpha)w_i + \alpha w} - b\left[(1-\beta)w_i + \beta w\right]p_i\right), \quad i = 1, 2, \ldots, n, \tag{15}$$

where p_i is the packet drop probability on the i-th path. Note that packets are dropped randomly with probabilities $p_1, \ldots, p_n$ rather than deterministically as in Assumption 1 of the previous section. Note also that the window size can be understood as the average on a coarse time scale, so (15) does not capture some detailed finer time scale behaviors.

Without loss of generality, we assume $p_1 \le p_2 \le \cdots \le p_n$. Setting the right-hand side of (15) to zero for the equilibria yields either $w_i^* = 0$ or

$$[(1-\alpha)w_i^* + \alpha w^*][(1-\beta)w_i^* + \beta w^*] = \frac{ab^{-1}(1-p_i)}{p_i}, \tag{16}$$

from which it follows that

$$p_i \le \frac{ab^{-1}}{\alpha\beta(w^*)^2 + ab^{-1}}, \quad \text{if } w_i^* > 0. \tag{17}$$

with equality if and only if $\alpha = \beta = 1$.

On the other hand, if $w_i^* = 0$, then for the equilibrium to be stable, it is necessary that the term in the parentheses of the right-hand side of (15) be non-positive, or, after rearrangement,

$$p_i \ge \frac{ab^{-1}}{\alpha\beta(w^*)^2 + ab^{-1}}, \quad \text{if } w_i^* = 0. \tag{18}$$

A comparison of (17) and (18) shows that in a stable equilibrium, there exists an N such that $w_i^* > 0$ if and only if $i \le N$, and

$$p_N \le \frac{ab^{-1}}{\alpha\beta(w^*)^2 + ab^{-1}} \le p_{N+1}, \tag{19}$$

where the first equality holds if and only if $\alpha = \beta = 1$.

We consider four cases for different values of α and β.

Case 1 $\alpha \ne 1$, $\beta = 1$. The equilibrium condition (16) gives

$$w_i^* = \frac{ab^{-1}(1-p_i)}{w^*(1-\alpha)p_i} - \frac{\alpha}{1-\alpha}w^*, \quad i \le N. \tag{20}$$

Summing over i and using $w_i^* = 0$ for $i > N$, we can solve for w^*,

$$w^* = \sqrt{\frac{ab^{-1}}{(1-\alpha)+\alpha N}\sum_{i=1}^{N}\frac{1-p_i}{p_i}}.$$

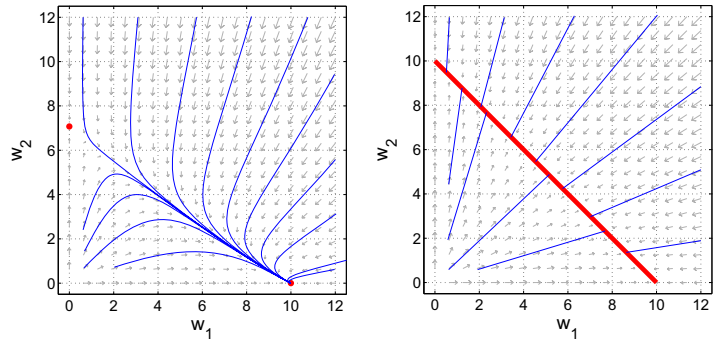
Substitution into (19) yields $N = \max\{k : 1 \le k \le n,\ H_k > \alpha\}$, where $H_k = \left(\frac{p_k}{1-p_k}\sum_{i=1}^{k}\frac{1-p_i}{p_i} - k + 1\right)^{-1}$ is monotonically decreasing in k. Depending on the value of α, a different subset of the paths is used. If $\alpha < H_n$, then $N = n$ and all the paths are used. If $\alpha \ge H_{K+1}$, where $K = \max\{i : p_i = p_1\}$ is the number of best paths, then $N = K$ and hence only the best paths are used. In particular, if $p_2 > p_1$ and $\alpha \ge \frac{p_1(1-p_2)}{p_2(1-p_1)}$, then $N = K = 1$, and only the first path is used.

Case 2 $\alpha = 1$, $\beta \ne 1$. Note that α and β play symmetric roles in (16) and (19), the analysis in Case 1 can be repeated verbatim with α replaced by β.

Case 3 $\alpha = 1$, $\beta = 1$. This is the limiting case of the previous two cases. The equilibrium condition (16) becomes

$$w^* = \sqrt{\frac{ab^{-1}(1-p_1)}{\alpha\beta p_1}} = \sqrt{\frac{ab^{-1}(1-p_i)}{\alpha\beta p_i}}, \quad \text{for } i \le N,$$

which is possible only if $p_i = p_1$ for $i \le N$. This is consistent with the results in Cases 1 and 2 when either $\alpha \to 1$ or $\beta \to 1$. However, in this case, any point on the simplex $\sum_{i=1}^{N} w_i^* = w^*$ is an equilibrium. If $p_1 < p_2$, then we have a unique stable equilibrium. If $p_1 = p_2$, then we have an uncountable number of equilibria, none of which are stable, although the simplex is a stable set. This is illustrated for $n = 2$ and $\tau_1 = \tau_2$ in Figure 7.



(a) $p_1 = 0.02$, $p_2 = 0.04$ (b) $p_1 = p_2 = 0.02$

Figure 7: Directional fields and trajectories for Fully Coupled MTCP. $\tau_1 = \tau_2 = 0.02$, $a = 1$, $b = 1/2$. The red dots and red lines represent the equilibria.

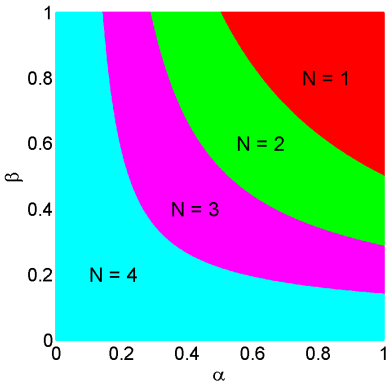


Figure 8: Number of paths used for different α and β. $n = 4$, $p_1 = 0.01$, $p_2 = 0.02$, $p_3 = 0.03$, $p_4 = 0.05$.

Figure 7(a) shows the directional field and some trajectories for $p_1 < p_2$. There are two equilibrium points,

$(\sqrt{2(1-p_1)/p_1}, 0)$ and $(0, \sqrt{2(1-p_2)/p_2})$. Only $(\sqrt{2(1-p_1)/p_1}, 0)$ is stable, and the trajectory converges to $(\sqrt{2(1-p_1)/p_1}, 0)$ except when $w_1 = 0$ initially.

Figure 7(b) shows the directional field and some trajectories for $p_1 = p_2$. There is a continuum of equilibrium points, all in the simplex $w_1 + w_2 = \sqrt{2(1-p_1)/p_1}$. All trajectories converge to the simplex as $t \to \infty$, but there is no preferred equilibrium and the limit depends on the initial condition.

Case 4 $\alpha \neq 1$ and $\beta \neq 1$. Note that for a given w^*, (16) has at most one positive root, so for $i \leq N$, we have

$$\frac{w_i^*}{w^*} = \frac{\sqrt{(\alpha-\beta)^2 + \frac{4ab^{-1}(1-\alpha)(1-\beta)(1-p_i)}{(w^*)^2 p_i}} - (\alpha + \beta - 2\alpha\beta)}{2(1-\alpha)(1-\beta)},$$

which shows that $w_i^* = w_j^*$ iff $p_i = p_j$. Summing over i from 1 to N yields

$$2(1-\alpha)(1-\beta) + N(\alpha + \beta - 2\alpha\beta) = \sum_{i=1}^{N} \sqrt{(\alpha-\beta)^2 + \frac{4ab^{-1}(1-\alpha)(1-\beta)(1-p_i)}{(w^*)^2 p_i}},$$

which, given N, has a unique positive solution. However, we do not have an explicit expression for w^*. Substitution of (19) into the above equation yields $N = \max\{k : 1 \leq k \leq n,\ G_k > 0\}$, where $G_k = 2(1-\alpha)(1-\beta) + k(\alpha + \beta - 2\alpha\beta) - \sum_{i=1}^{k} \sqrt{(\alpha-\beta)^2 + 4\alpha\beta(1-\alpha)(1-\beta)p_k(1-p_k)^{-1}(1-p_i)p_i^{-1}}$.

Given the packet drop probabilities p_k, the unit square in the α-β plane is then partitioned into different regions according to different values of N, as illustrated in Figure 8. When the coupling is strong in both the increase and decrease phases, traffic is sent over only the best paths. As the coupling gets weaker in either phase, more non-best paths are used as well, with better paths carrying more traffic. In particular, paths with the same quality, i.e. loss probability, have the same amount of traffic, except for the extreme case where both phases have the strongest coupling ($\alpha = \beta = 1$).

4.2 Flappiness

In this subsection, we look at the short-term traffic distribution among different paths, focusing on the so-called *flappy* behavior, namely the traffic of a flow tends to concentrate on one path and then another [18, 16]. For simplicity, we start with the FULLY COUPLED case, i.e. $\alpha = \beta = 1$, for $n = 2$ subflows. To mimic TCP Reno, we set $a = 1$ and $b = 1/2$. We consider two scenarios, $|w_1 - w_2| \gg 1$ and $w_1 \approx w_2$.

Scenario 1 $|w_1 - w_2| \gg 1$. Without loss of generality, assume $w_1 \gg w_2$. When there is no loss,

$$\begin{cases} \dfrac{d}{dt} w_1 = \dfrac{1}{\tau_1} \dfrac{w_1}{w_1 + w_2} \approx \dfrac{1}{\tau_1}, \\ \dfrac{d}{dt} w_2 = \dfrac{1}{\tau_2} \dfrac{w_2}{w_1 + w_2} \approx \dfrac{1}{\tau_2} \dfrac{w_2}{w_1} \ll \dfrac{1}{\tau_2}. \end{cases}$$

When subflow 1 suffers a loss, its window size becomes

$$(w_1)_{new} = w_1 - \frac{w_1 + w_2}{2} = \frac{w_1 - w_2}{2} \approx \frac{w_1}{2} \gg w_2.$$

When subflow 2 suffers a loss, its window size becomes

$$(w_2)_{new} = \max\left\{w_2 - \frac{w_1 + w_2}{2}, 1\right\} = 1 \ll w_1.$$

Note the inhibition effect. The dominating subflow behaves almost like ordinary TCP, while the dominated one is almost stifled: its window size increases much more slowly between losses than that of ordinary TCP. Moreover, a single loss simply reduces its whole effort to nil. Therefore, when this scenario occurs, it will continue for a long time until the following happens: there is a long period of time in which subflow 2 has no loss and subflow 1 has a batch of losses, as a result of which, $w_1 \approx w_2$ and we are in Scenario 2.

Scenario 2 $w_1 \approx w_2 \gg 1$. In this scenario, the two subflows compete with each other on almost equal footing. Whichever subflow suffers a loss first loses the battle. Assume subflow 1 suffers a loss. Then its window size becomes

$$(w_1)_{new} = \max\left\{w_1 - \frac{w_1 + w_2}{2}, 1\right\} \approx 1 \ll w_2.$$

The pressure on subflow 2 is now released and we return to Scenario 1 again. Note that a single loss from either subflow will produce such a transition to Scenario 1, so Scenario 2 does not last long and the flow spends most of its time in Scenario 1.

If $p_1 < p_2$, subflow 1 has a better chance to win in Scenario 2, so it is more likely to transition to $w_1 \gg w_2$. It is also more difficult to transition to the scenario $w_1 \approx w_2$ from $w_1 \gg w_2$ than from $w_2 \gg w_1$, so we have $w_1 \gg w_2$ most of the time. This is consistent with the analysis in Section 4.1 that $(\sqrt{2(1-p_1)/p_1}, 0)$ is the only stable equilibrium of (15) when $p_1 < p_2$.

If $p_1 = p_2$, both subflows are equally likely to win in Scenario 2, so it is equally likely to transition to $w_1 \gg w_2$ or $w_2 \gg w_1$. This is also consistent with the equilibrium analysis in the sense that both $(\sqrt{2(1-p_1)/p_1}, 0)$ and $(0, \sqrt{2(1-p_2)/p_2})$ are unstable equilibria of (15). However, the equilibrium analysis does not predict the fact that the flow spends very little time with $w_1 \approx w_2$, though (15) also has many equilibria with $w_1 \approx w_2$. This is because the observed behavior works on a finer time scale than is captured by (15); see comments following (15).

Note that flappiness is associated with homogenous paths. We will henceforth assume the paths are almost homogenous by focusing on the good paths, namely those paths whose packet drop probabilities are close to the lowest. The addition of other paths does not affect the behavior significantly.

The previous analysis naturally extends to the cases where $n \geq 3$. If there is a subflow that dominates the rest, then it dominates for a long time. However, since there are many subflows competing with each other, it is more difficult to establish the dominance, so the flappy behavior is observed less often.

Consider now the impact of b on flappiness. Suppose we decrease b. If the flow is in Scenario 2, it will take more than a single loss event to move it to Scenario 1, so the flow will remain longer around $w_1 \approx w_2$. On the other hand, if it is in Scenario 1, the window decrease incurred through a loss event gets smaller for the larger subflow and remains unchanged for the smaller subflow, so it is more difficult to return to Scenario 2. Thus decreasing b makes the transitions between the two scenarios slower and less frequent.

Now suppose we increase b. The transition from Scenario 2 to Scenario 1 is almost unaffected, so the flow does not stay long around $w_1 \approx w_2$. For Scenario 1, the window decrease incurred through a loss event for the larger subflow becomes larger and remains unchanged for the smaller subflow, so it is easier to return to Scenario 2. Thus the flips between the two extremes are more frequent.

What are the impacts of α and β? We will fix $b = 1/2$, $n = 2$ and consider the two extreme cases, i.e. Coupled Increase and Coupled Decrease. Consider Coupled Increase first. In Scenario 2, the two subflows behave almost like independent TCP flows, though the windows increase at a smaller rate, so they can stay around $w_1 \approx w_2$ for a long time before transitioning to Scenario 1. In Scenario 1, the smaller subflow increases at a small rate due to the inhibition effect, but the penalty upon a loss event is also small, so it does not take long to return to Scenario 2. Therefore, there is no prominent flappy behavior.

Now consider Coupled Decrease. The analysis for Scenario 2 is unaffected. For Scenario 1, there is no inhibition effect, so the smaller subflow can increase more freely between two consecutive losses, and its window size tends to be larger. Consequently, the window decrease incurred through a loss event for the larger subflow is also larger. Thus the domination of the larger subflow is weaker, making transitions to Scenario 2 easier, and we do not observe long periods during which essentially only one subflow has positive rate.

To summarize, the flappy behavior is prominent only when there is strong coupling in both the increase and decrease phases, and when the number of good paths is small (but more than one).

5 Experimental Evaluation

In this section we verify the theoretical results in Sections 3 and 4, and extend our study of the MTC algorithms via simulation, accounting for the large buffer case. In particular we show: (1) how the link utilization is affected by the number of subflows, the buffer size of bottleneck links, and the number of co-existing MTC flows; (2) how different MTC algorithms dynamically distribute traffic over the subflows.

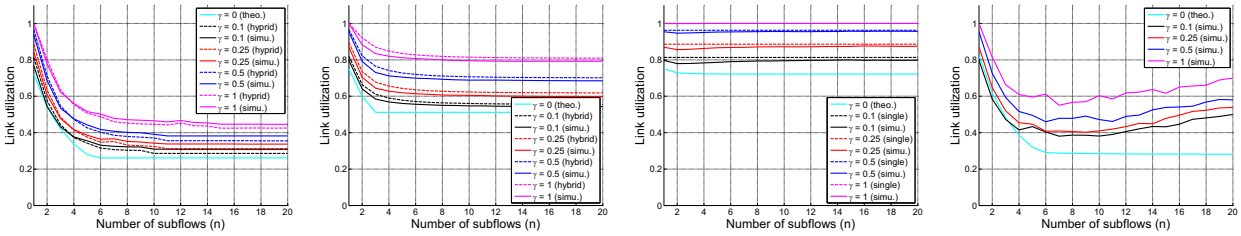
The topology used in the experiments is similar to that in Figure 1. There are n distinct paths between

nodes S and R. There are multiple co-existing MTC flows, each consisting of n subflows, with the i-th subflow going over the i-th path. All n paths have the same link capacity, propagation delay, and bottleneck buffer size. The particular parameter settings vary in different experiments.

5.1 Link Utilization

In this subsection, we evaluate the link utilization of MTC flows in the absence of background traffic. This corresponds to a situation in which the network bandwidth is shared solely among MTC applications.

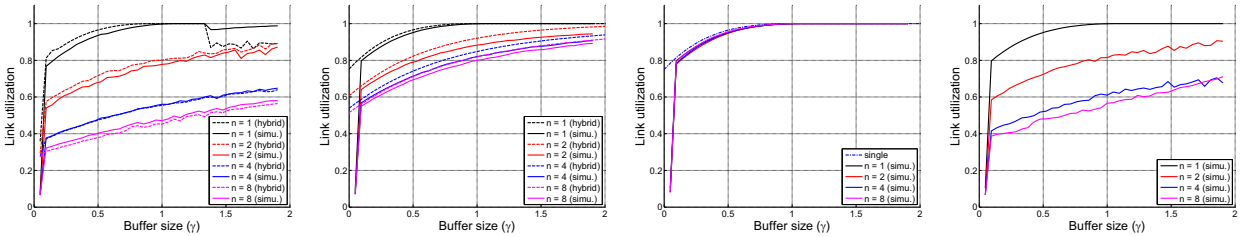
In the first two sets of experiments, there is only one MTC flow between nodes S and R. The size of each data packet is 1.5KBytes. Each path has a bottleneck link capacity of 5Mbps and a propagation delay of 100ms. Thus the capacity-delay product is roughly equal to 0.5Mbits (approximately 42pkts). The buffer size at the bottleneck link varies from 2pkts to 84pkts, corresponding to values of γ from 0.05 to 2. Each run of the experiments lasts 1200s and steady state is reached by 200s.



(a) MSTCP (b) Coupled Decrease MTCP (c) Coupled Increase MTCP (d) Fully Coupled MTCP

Figure 9: Link utilization vs. number of subflows

Figure 9 plots the link utilization versus the number of subflows n, parameterized by the buffer size at the bottleneck link. The dashed curves in Figure 9(a) and 9(b) are calculated using (6) and (10), respectively, where f is measured from experiments. The dashed curves in Figure 9(c) correspond to the theoretical values of a single path TCP. Note that the dashed curves match pretty well the solid curves from simulations, thus verifying the theoretical results. Now we make the following observations. First, the link utilization is lower bounded by a constant in all four cases. Second, for Coupled Increase MTCP, the number of subflows has very little impact on the link utilization, whereas in the other three cases, a larger number of subflows leads to a lower link utilization. Third, for Fully Coupled MTCP, as the number of subflows increases, the link utilization first decreases rapidly and then starts to increase slowly. It is not monotonically decreasing as we would expect from the analysis in Section 3.2.3. This is because Assumption 2 does not hold and some subflows can have higher throughput than predicted there. If we take a closer look at Coupled Increase MTCP, similar non-monotonicity is observed and the same explanation applies. For the MTC algorithms other than Fully Coupled MTCP, the link utilization approaches a limit after the number of subflows exceeds four.

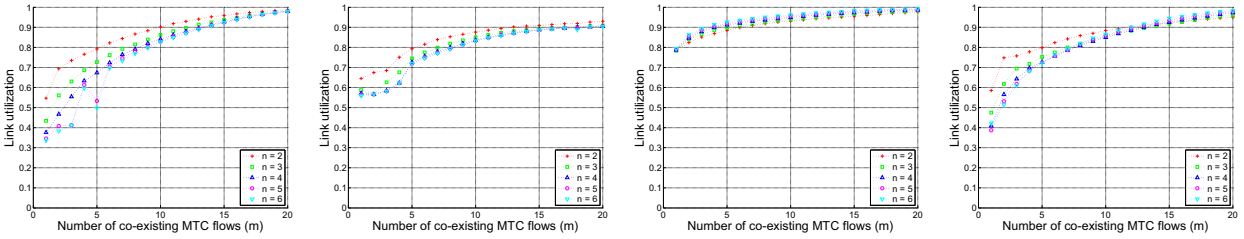


(a) MSTCP (b) Coupled Decrease MTCP (c) Coupled Increase MTCP (d) Fully Coupled MTCP

Figure 10: Link utilization vs. bottleneck buffer size

Figure 10 plots link utilization versus the buffer size of the bottleneck link, parameterized by the number of subflows. The dashed curves are calculated as in Figure 9, and match well the solid curves from simulations, except when $n = 1$ and $\gamma > 1.3$ in Figure 10(a). The discrepancy there is attributed to the invalidness of the model in Section 3.1 in the presence of time-out events. For all four cases, the link utilization increases as the buffer size of the bottleneck link increases. This is consistent with the fact that a larger buffer can help reduce packet drops, thus increasing the throughput of the underlying TCP flows.

The parameters in the next set of experiments are as follows. There are multiple co-existing MTC flows between S and R. The size of each data packet is 1KBytes. Each path has a bottleneck link capacity of 10Mbps and a propagation delay of 80ms. Thus the capacity-delay product is roughly equal to 0.8Mbits (i.e., 100pkts). The buffer size at the bottleneck link is 5pkts. Each simulation runs for 1200s and steady state is reached by 200s.



(a) MSTCP (b) Coupled Decrease MTCP (c) Coupled Increase MTCP (d) Fully Coupled MTCP

Figure 11: Link utilization vs. number of co-existing MTC flows

Figure 11 plots link utilization versus the number m of co-existing MTC flows. We make the following observations. First, in all four cases link utilization increases as the number of co-existing MTC flows increases and the curves resemble those in Figure 6. As discussed in Section 3.3, the increasing link utilization is due to better statistical multiplexing. Second, for Coupled Increase MTCP, the number of subflows has little impact on link utilization, whereas in the other three cases, a larger number of subflows leads to a lower link utilization. Third, the link utilization approaches a limit after the number of subflows exceeds four. These observations generalize the corresponding observations made for a single flow (Figure 9) to the multi-flow

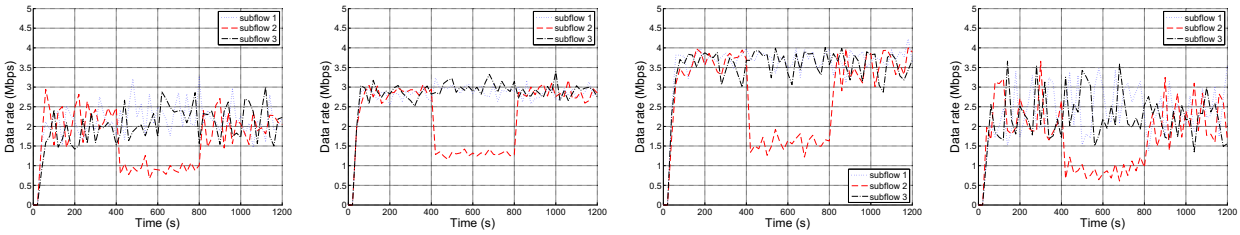
case. Fourth, when the number of co-existing MTC flows becomes very large, link utilization approaches a limit independent of the number of subflows, and this limit is one except for Coupled Decrease MTCP.

5.2 Traffic Distribution

The theoretical analysis in Section 4.1 shows that different coupling strengths lead to the usage of different subsets of paths in the long run. In particular, according to the theory, Fully Coupled MTCP uses only the best path(s). In this subsection we complement the theoretical analysis with simulations.

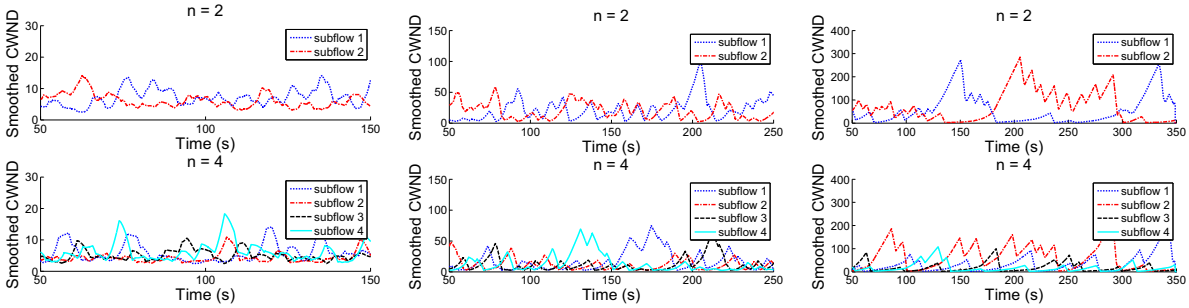
The experimental setup is as follows. There is only one MTC flow between nodes S and R, which consists of three subflows. Each data packet is 1KBytes. Each path has a bottleneck link capacity of 10Mbps and a propagation delay of 80ms. The buffer size of the bottleneck link is 20pkts. An aggregate background UDP traffic is injected into each path to introduce randomness. The background UDP traffic on each path is the sum of 40 Markov On-Off flows, each having a peak rate of 0.25Mbps, an average On period of 100ms and an average Off period of 100ms. This reduces the average capacity available to MTC flows to 5Mbps. In addition, subflow 2 experiences a surge of UDP traffic in the interval between 400s and 800s, during which an additional aggregate of 20 UDP flows with the same parameters is added to path 2. This further reduces the capacity available to the MTC flow on that path to 2.5Mbps.

Figure 12 plots the average data rate over 20 second intervals. The 20 second interval has been chosen to give enough but not too much smoothing for easy visualization. These plots illustrate how the four MTC algorithms shift traffic among subflows in response to changes in the congestion level. Except at the beginning and in the small intervals around 400s and 800s, the flow is in steady state. As expected, during the period of 400s-800s, the subflow on the more congested path has a lower rate than the subflows on the less congested paths. However, rather than use the best paths only as discussed in Section 4.1, Fully Coupled MTCP places an appreciable amount of traffic on the more congested path (see Figure 12(d)). This is because the fluid model is not a faithful representation of the window-based mechanism. In contrast to the fluid model, decrements are discrete jumps, and the window on the more congested path, though it drops more frequently, does increase between decrements, thus resulting in a nonnegligible throughput. Among the three AIMD MTC algorithms, Coupled Increase MTCP achieves the highest throughput while Fully Coupled MTCP achieves the lowest. This is consistent with the theoretical analysis in Section 3 and the simulation results in the previous subsection. Note also that Fully Coupled MTCP and MSTCP have similar instantaneous rates, which oscillate more dramatically than Coupled Increase and Coupled Decrease MTCP.



(a) MSTCP (b) Coupled Decrease MTCP (c) Coupled Increase MTCP (d) Fully Coupled MTCP

Figure 12: Average data rate over 20 second intervals.



(a) A buffer size of 4pkts (b) A buffer size of 10pkts (c) A buffer size of 20pkts

Figure 13: Trajectories of the congestion windows of Fully Coupled MTCP.

5.3 Flappiness

In this subsection we investigate the flappy behavior of the MTC algorithms. The experimental setup is similar to that of Section 5.2, but the background traffic is from real Internet traffic traces captured at an edge router of the UMass campus network. The traffic traces are chosen such that in each simulation run, they have similar statistics on all paths in terms of long-term throughput and packet drop probability. The capacity of each path is 100Mbps and the buffer size takes values of 4, 10, and 20pkts, resulting in a packet drop probability on each path of around 4.2%, 0.5%, and 0.05%, respectively. The different packet drop probabilities then lead to different average window sizes.

Figure 13 shows the sample trajectories of the congestion windows of Fully Coupled MTCP flows with 2 and 4 subflows, respectively. For easy visualization, the trajectories are smoothed using a central moving average within a window of 20s. Note that in the top plot of Figure 13(c), most of the time one subflow dominates the other, and the domination alternates between them. This is an example of flappy behavior. A comparison of the top plots in all three panels shows that the larger the average window size, the more prominent the flappy behavior, since a large window size helps maintain the dominance. The bottom plots show a similar trend. On the other hand, a comparison of the two plots within each panel shows that a larger number of subflows result in less prominent flappiness. With 4 subflows, there is a dominant subflow from time to time, but this occurs less often and is less prominent.

We also conducted simulations for the other MTC algorithms and found similar trends. The detailed experimental results are omitted. We mention, however, under the same conditions, Coupled Increase and Coupled Decrease MTCPs exhibit less prominent flappy behavior than Fully Coupled MTCP, and the periods during which one subflow dominates are much shorter for MSTCP than for Fully Coupled MTCP.

6 Related Work

In recent years there has been a great deal of interest in multipath transmission control (MTC) [17, 5, 6, 9, 18, 19]. Kelly et al. [7] established a framework in which one can show that the TCP congestion control algorithm solves a network resource optimization problem, the optimum of which achieves the fairness corresponding to the congestion control algorithm. Based on this work, two research groups, Han et al. and Kelly et al., proposed two rate-based joint congestion controllers and analyzed the stabilities of the controllers in [5] and [6], respectively.

Following these theoretical works, a lot of effort has been made towards exploring the possibility of incorporating MTC mechanisms into the current Internet protocol stack [17, 11, 14]. Kokku et al. [11] proposed a multipath-based background transfer system named Harp in which an MTC controller based on Scalable TCP [8] was used. Simulation in ns2 and experiments in planet-lab [15] show that, compared to its counterparts based on ordinary TCP, Harp achieves higher throughput and alleviate local congestions. Mallada and Paganini demonstrated the feasibility of running MTCP over the Internet by incorporating multipath support to the current TCP-FAST and Routing Information Protocol (RIP) [14].

Wischik et al. [18, 19] implemented the joint congestion controllers proposed in [5] and [6] and observed the flappy behavior of these controllers. A preliminary analysis was given and an empirical solution provided. Our work on flappiness differs from theirs in that we study flappiness in the context of a family of additive-increase and multiplicative-decrease MTC algorithms and give a more comprehensive analysis of different impacts of various parameters.

7 Conclusions

In this paper, we study four variations of multipath transmission control algorithms, focusing on the performance issues uncovered by recent experimental studies: low link utilization when the number of simultaneous flows is small, and flappiness. We provide a quantitative explanation for the observed low link utilization. Our analysis reveals the different natures of the couplings in the increase and decrease phases of multipath transmission control algorithms. The coupling in the increase phase reduces the link utilization by changing the window trajectory from linear to convex, while the coupling in the decrease phase simply incurs a larger

decrement of the window size upon a loss event. We also provide a systematic analysis for the traffic distribution among different subflows of an multipath transmission control flow. The equilibrium analysis shows that, depending on the strength of couplings, the traffic is distributed differently among paths of different loss probabilities in the long term. The impact of different parameters on the short-term flappy behavior are also analyzed. It is found that the flappy behavior is prominent only when there is strong coupling in both the increase and decrease phases, and when the number of good paths is small. However, the notion of flappiness used here is not precise and appeals to our intuition. We intend to formalize this notion with a quantitative measure in future work.

Acknowledgement This work is supported in part by the National Science Foundation under grants IIS-0916726, CNS-1040781, CNS-0721790 and EFRI-0735974, and by ARO MURI under grant W911NF-08-1-0233.

References

[1] M. Allman, V. Paxson, and W. Stevens. RFC 3439: TCP Congestion Control. Apr. 1999.

[2] J.-Y. Boudec. Rate adaptation, congestion control and fairness: A tutorial, 2008. http://ica1www.epfl.ch/PS_files/LEB3132.pdf.

[3] B. Cohen. Incentives built robustness in BitTorrent. In *Proc. of P2P Economics workshop*, 2003.

[4] Firefox Addon: DownThemAll. http://www.downthemall.net/.

[5] H. Han, S. Shakkottai, C. V. Hollot, R. Srikant, and D. Towsley. Multi-path TCP: a joint congestion control and routing scheme to exploit path diversity in the Internet. *IEEE/ACM Transactions on Networking (TON)*, 14(6):1260–1271, Dec. 2006.

[6] F. Kelly and T. Voice. Stability of end-to-end algorithms for joint routing and rate control. *ACM SIGCOMM Computer Communication Review*, 35(2):5–12, Apr. 2005.

[7] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan. Rate control for communication networks: Shadow prices, proportional fairness and stability. *The Journal of the Operational Research Society*, 49(3):237–252, Mar. 1998.

[8] T. Kelly. Scalable TCP: Improving performance in highspeed wide area networks. *ACM SIGCOMM Computer Communication Review*, 32(2):83–91, Apr. 2003.

[9] P. Key, L. Massoulié, , and D. Towsley. Combining multipath routing and congestion control for robustness. In *Proc. of 40th Conference on Information Sciences and Systems (CISS)*, 2006.

[10] P. B. Key, L. Massoulié, and D. F. Towsley. Path selection and multipath congestion control. *Commun. ACM*, 54(1):109–116, 2011.

[11] R. Kokku, A. Bohra, S. Ganguly, and A. Venkataramani. A multipath background network architecture. In *Proc. of INFOCOM '07*, pages 1352–1360, 2007.

[12] P. P. C. Lee and V. M. D. Rubenstein. Distributed algorithms for secure multipath routing in attack-resistant networks. *IEEE/ACM Transactions on Networking (TON)*, 15(6):1490–1501, Dec. 2007.

[13] W. Lou, W. Liu, Y. Zhang, and Y. Fang. Spread: Improving network security by multipath routing in mobile ad hoc networks. *Wireless Networks*, 15(3):279–294, Apr. 2009.

[14] E. Mallada and F. Paganini. Optimal congestion control with multipath routing using TCP-FAST and a variant of RIP. *Lecture Notes In Computer Science*, pages 205–214, 2007.

[15] PlanetLab. http://www.planet-lab.org/.

[16] C. Raiciu, D. Wischik, and M. Handley. Practical congestion control for multipath transport protocols. Technical report, University College London, 2010.

[17] W.-H. Wang, M. Palaniswami, and S. H. Low. Optimal flow control and routing in multi-path networks. *Performance Evaluation*, 52(2):119–132, 2003.

[18] D. Wischik, M. Handley, and C. Raiciu. Control of multipath TCP and optimization of multipath routing in the Internet. *Lecture Notes In Computer Science*, pages 204–218, 2009.

[19] D. Wischik, C. Raiciu, A. Greenhalgh, and M. Handley. Design, implementation and evaluation of congestion control for multipath TCP. In *Proc. of NSDI'11*, 2011.

Appendices

A Proof of (8)

Under Assumption 3, (5) yields

$$\frac{dw_{i-1}}{w_{i-1}} = \frac{dw_i}{w_i}. \tag{21}$$

Integrating (21) over t on $(0, L/n]$ yields

$$\frac{w_i(0^+)}{w_{i-1}(0^+)} = \frac{w_i(L/n)}{w_{i-1}(L/n)}, \tag{22}$$

where $i-1$ is understood to be n when $i = 1$. From (7), we have

$$w_i(L/n) = w_{i-1}(0). \tag{23}$$

Substituting (23) into (22) yields

$$\frac{w_1(0^+)}{w_n(0)} = \frac{w_n(0)}{w_{n-1}(0)} = \cdots = \frac{w_3(0)}{w_2(0)} = \frac{w_2(0)}{w_1(0)}.$$

By multiplying all the terms, we get the common ratio of the geometric progression $\{w_i(0)\}$,

$$\frac{w_{i+1}(0)}{w_i(0)} = \left(\frac{w_1(0^+)}{w_1(0)}\right)^{1/n} = f^{1/n}, \quad i = 1, 2, \ldots, n-1,$$

where we have used $w_1(0) = W_{\max}$ and $w_1(0^+) = fW_{\max}$ given by ASSUMPTION 2. It then follows that

$$w_i(0) = w_1(0) f^{\frac{i-1}{n}}, \quad i = 1, 2, \ldots, n, \tag{24}$$

and hence

$$w(0) = \sum_{i=1}^{n} w_i(0) = w_1(0) \frac{1-f}{1-f^{1/n}}. \tag{25}$$

According to (1), we have

$$w_1(0^+) = \max\{w_1(0) - bw(0), 1\}, \tag{26}$$

which, upon using (24), gives (8).

B Proofs of (10) and (11)

As in the analysis for MSTCP in Section 3.1, suppose $w_i(0) = W_{\max}$ and let $w_i(0^+) = fW_{\max}$ be the window size immediately after the decrease. If $w_i(0^+) \le W_0$, then the window size will grow linearly until it reaches W_0 at time t_0, where

$$t_0 = \frac{D}{a}[W_0 - w_i(0^+)] = \frac{W_0}{aC}[W_0 - w_i(0^+)].$$

It then grows sublinearly until it reaches $W_{\max}$ at time t_1, where

$$t_1 = t_0 + \frac{1}{2aC}(W_{\max}^2 - W_0^2) = \frac{W_0^2}{2aC}(\kappa^2 + 1 - 2\kappa f). \tag{27}$$

By (4) and (9), the data rate is $r_i(t) = a^{-1} w_i dw_i/dt$. Using (27), the throughput of the i-th path in the given cycle is

$$T_{\text{cycle}} = \frac{1}{t_1}\int_0^{t_1} \frac{1}{2a}\frac{dw_i^2}{dt}dt = C\frac{\kappa^2(1-f^2)}{\kappa^2 + 1 - 2\kappa f}.$$

If $w_i(0^+) > W_0$, then $T_{\text{cycle}} = C$. Therefore, the link utilization of a cycle with initial window size $fW_{\max}$ is

$$U_{\text{cycle}} = \frac{T_{\text{cycle}}}{C} = \begin{cases} \dfrac{\kappa^2(1-f^2)}{\kappa^2 + 1 - 2\kappa f}, & \kappa f \le 1, \\ 1, & \kappa f \ge 1. \end{cases} \tag{10}$$

Next we prove (11). Under ASSUMPTION 3, the solution to (9) is

$$w_i(t) = w_i(0^+) + \frac{t}{L}[w_i(L) - w_i(0^+)], \quad 0 < t \le L. \tag{28}$$

Using (7),

$$w(0) = \sum_{i=1}^{n} w_1\left(\frac{i}{n}L\right) = \frac{n+1}{2}w_1(0) + \frac{n-1}{2}w_1(0^+).$$

Substitution into (26) then yields

$$f = \max\left\{\frac{2-b(n+1)}{2+b(n-1)}, \frac{1}{W_0}\right\}. \tag{11}$$

C Proof of (12)

Between two consecutive losses of the COUPLED INCREASE MTCP flow, which come from two different paths, the window dynamics are described by the following system of ODEs,

$$\frac{d}{dt}w_i = \frac{aw_i}{\tau_i w} = \frac{aw_i}{Dw}, \quad i = 1, 2, \ldots, n. \tag{29}$$

Note that (29) implies (21), so the derivation there carries over. Using (24), we have

$$w(0^+) = w_1(0^+) + \sum_{i=2}^{n} w_i(0) = w_1(0)\frac{f^{1/n}(1-f)}{1-f^{1/n}}.$$

Since $w(t)$ has period L/n, (25) yields

$$w(L/n) = w(0) = w_1(0)\frac{1-f}{1-f^{1/n}}.$$

Summing over i in (29), we have

$$\frac{d}{dt}w = \frac{a}{D}, \tag{30}$$

so $w(t)$ is linear on $(0, L/n]$. Therefore, the throughput of the flow during the period $(0, L/n]$ is

$$T_{\text{tot}} = \frac{w(0^+) + w(L/n)}{2D} = \frac{W_{\max}}{2D}\frac{(1-f)(1+f^{1/n})}{1-f^{1/n}}.$$

The link utilization is then

$$U = \frac{T_{\text{tot}}}{nC} = \frac{\kappa(1-f)(1+f^{1/n})}{2n(1-f^{1/n})}. \tag{12}$$

Since $w_1(0^+) = \max\{(1-b)w_1(0), 1\}$, we have

$$f = \max\{1-b, W_0^{-1}\}.$$

D Proofs of the Formulas in Table 1

By ASSUMPTION 4, there exists a window size $W_{\text{eff}} = w_1^1(0)$ that plays the role of W_0 in the single flow case, i.e. a loss occurs on the i-th path for the j-th flow if and only if $w_i^j(t)$ exceeds W_{eff}. If we know W_{eff}, then conceptually we can decouple the m flows from each other and think of them as using different sets of paths, each with a capacity-delay product W_{eff}. The overall link utilization is then given by

$$U = \frac{mW_{\text{eff}}}{W_0} U_{\text{single}}, \tag{31}$$

where U_{single} is the corresponding single flow link utilization in Sections 3.1 and 3.2 with W_0 replaced by W_{eff}. It now remains to find W_{eff}.

D.1 MSTCP

Consider the congestion windows $\{w_1^j(t)\}$ on the first path. By the same derivation leading to (24), we have

$$w_1^j(0) = W_{\text{eff}} f^{\frac{j-1}{m}}.$$

Since a loss occurs on the first path at time 0,

$$W_0 = \sum_{j=1}^{m} w_1^j(0) = W_{\text{eff}} \frac{1-f}{1-f^{1/m}},$$

and hence

$$W_{\text{eff}} = W_0 \frac{1-f^{1/m}}{1-f}. \tag{32}$$

Substitution into (31) and (6) yields

$$U = \frac{m(1-f^{1/m})}{-\log f}.$$

By (8), $f = \max\{(1-b)^n, W_{\text{eff}}^{-1}\}$. Using (32),

$$f = \max\{(1-b)^n, f^*\}, \tag{33}$$

where $f^* \in (0,1)$ is the root to

$$f^{1+1/m} - \left(1 + \frac{1}{W_0}\right) f + \frac{1}{W_0} = 0. \tag{34}$$

D.2 Coupled Decrease MTCP

By (28),

$$w_1^1(t) = w_1^1(0^+) + \frac{t}{L}[w_1^1(L) - w_1^1(0^+)] = \left[f + \frac{t}{L}(1-f)\right] W_{\text{eff}}, \quad 0 < t \le L.$$

Using (14),

$$W_0 = \sum_{j=1}^{m} w_1^j(0) = \sum_{j=1}^{m} w_1^1\left(\frac{j}{m}L\right) = \left(\frac{m+1}{2} + \frac{m-1}{2}f\right) W_{\text{eff}},$$

and hence

$$W_{\text{eff}} = \frac{2}{m+1+(m-1)f} W_0. \tag{35}$$

Substitution into (31) and (10) with $\kappa = 1$ yields

$$U = \frac{m(1+f)}{m+1+(m-1)f}.$$

By (11),

$$f = \max\left\{\frac{2-b(n+1)}{2+b(n-1)}, \frac{1}{W_{\text{eff}}}\right\},$$

which, combined with (35), yields

$$f = \max\left\{\frac{2-b(n+1)}{2+b(n-1)}, \frac{m+1}{2W_0 - m + 1}\right\}.$$

D.3 Coupled Increase MTCP

As before, we have by (14),

$$W_0 = \sum_{j=1}^{m} w_1^j(0) = \sum_{j=1}^{m} w_1^1\left(L - \frac{j-1}{m}L\right) = \sum_{j=1}^{m} w_1^1\left(\frac{j}{m}L\right). \tag{36}$$

Now we need to find the expression for $w_1^1(t)$. By (25) and (7),

$$w_1^1\left(\frac{n-i+1}{n}L\right) = w_i^1(0) = W_{\text{eff}} f^{-\frac{i-1}{n}}, \quad i = 1, 2, \ldots, n.$$

Thus

$$w_1^1\left(\frac{k}{n}L\right) = W_{\text{eff}} f^{1-k/n}, \quad k = 1, 2, \ldots, n. \tag{37}$$

By (22), $w_1^1(t)$ is a constant fraction of $w^1(t) = \sum_{i=1}^n w_i^1(t)$ on each interval $(kL/n, (k+1)L/n]$. Since (30) implies that $w^1(t)$ increases linearly on each such interval, $w_1^1(t)$ also increase linearly. Thus $w_1^1(t)$ is a broken line inscribed in the exponential curve $y(t) = W_{\text{eff}} f^{1-t/L}$ and hence

$$w_1^1(t) = \frac{n}{L}\left[\left(t - \frac{k}{n}L\right) w_1^1\left(\frac{k+1}{n}L\right) + \left(\frac{k+1}{n}L - t\right) w_1^1\left(\frac{k}{n}L + 0\right)\right], \quad \text{for } \frac{k}{n}L < t \le \frac{k+1}{n}L.$$

Using (37),

$$w_1^1\left(\frac{j}{m}L\right) = W_{\text{eff}} f^{1-\frac{1}{n}\lceil\frac{jn}{m}\rceil}\left[1 - \left(\left\lceil\frac{jn}{m}\right\rceil - \frac{jn}{m}\right)\left(1 - f^{1/n}\right)\right].$$

Plugging into (36),

$$W_0 = g(f, n, m) W_{\text{eff}}, \tag{38}$$

where

$$g(f, n, m) = \sum_{j=1}^{m} f^{1-\frac{1}{n}\lceil\frac{jn}{m}\rceil}\left[1 - \left(\left\lceil\frac{jn}{m}\right\rceil - \frac{jn}{m}\right)\left(1 - f^{1/n}\right)\right].$$

Substitution of (38) into (31) and (12) then yields,

$$U = \frac{m(1-f)(1+f^{1/n})}{2n(1-f^{1/n})g(f, n, m)}. \tag{39}$$

Using $f = \max\{1-b, W_{\text{eff}}^{-1}\}$ from Section 3.2.2, f is given by

$$f = \max\{1-b, f^{**}\},$$

where $f^{**} \in (0,1)$ is the root to

$$g(f, n, m) - fW_0 = 0. \tag{40}$$

D.4 Fully Coupled MTCP

Subsitution of (38) into (31), (12) and (8) shows that the overall link utilization is given by (39) with $f = \max\{(1-b)^n, f^{**}\}$, where f^{**} is the root to (40).