Network Congestion Control with Markovian Multipath Routing

Roberto Cominetti
Departamento de Ingenieria Industrial
Universidad de Chile
Santiago, Chile
Email: rccc@dii.uchile.cl

Cristóbal Guzmán
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: cguzman@gatech.edu

***Abstract*—In this paper we consider an integrated model for TCP/IP protocols with multipath routing. The model combines a Network Utility Maximization for rate control based on end-to-end queuing delays, with a Markovian Traffic Equilibrium for routing based on total expected delays. We prove the existence of a unique equilibrium state which is characterized as the solution of an unconstrained strictly convex program. A distributed algorithm for solving this optimization problem is proposed, with a brief discussion of how it can be implemented by adapting the current Internet protocols.**

I. Introduction

Routing and congestion control are basic components of packet-switched communication networks. While *routing* is responsible for determining efficient paths along which the sources communicate to their corresponding receivers, *congestion control* manages the transmission rate of each source in order to keep network congestion within reasonable limits. In current practice both mechanisms belong to separate design layers that operate on different time-scales: the IP layer (Internet Protocol) determines single-path routings which are updated on a slow time-scale, while the TCP layer (Transmission Control Protocol) corresponds to end-to-end users that perform rate congestion control at a faster pace for which routing can be considered to be fixed. Scalability considerations impose that these protocols must operate in a decentralized manner.

Roughly speaking, TCP controls the rate of a source by managing a *window size* that bounds the maximum number of outstanding packets that have been transmitted but not yet acknowledged by the receiver. Once this window size is reached the source must wait for an acknowledgment before sending a new packet, so the rate is approximately one window of packets per round-trip time. As the network gets congested, the round-trip time increases and the transmission rate is automatically slowed down. In addition, TCP dynamically adjusts the window size of a source in response to network congestion. To this end, links generate a scalar measure of their own congestion (e.g. packet loss probability, average queue length, queueing delay) and each source is fed back a *congestion signal* that reflects the aggregate congestion of the links along its route. This signal is used by the source to adjust its window size so that the higher the congestion the smaller the rate. The predominant TCP protocols in use are Tahoe and Reno which use *packet loss* as congestion measure, and Vegas which is based on *queuing delay*. We refer to [1] for a description and comparison of current protocols and their models.

The interaction of many sources performing a decentralized congestion control based on feedback signals that are subject to estimation errors and communication delays, gives rise to complex dynamics that are difficult to analyze. However, assuming that the dynamics stabilize on a steady state, the equilibrium can be characterized as an optimal solution of a Network Utility Maximization (NUM) problem (see [2], [3], [4]). Thus, the TCP mechanism can be viewed as a decentralized algorithm that seeks to optimize an aggregate utility function subject to network constraints. The NUM approach also allows to compare different protocols in regard to their fairness and efficiency.

Partially supported by FONDECYT 1100046 and Instituto Milenio Sistemas Complejos de Ingeniería.

This research was done while the second author was a student of Departmento de Ingeniería Matemática, Universidad de Chile.

A second element of packet-switched networks is *routing*. This function is performed by routers in a decentralized manner using routing tables that determine the next hop for each destination. The routing tables are updated periodically by an asynchronous Dijkstra-type iteration that computes optimal paths according to some metric such as hop count, latency, delay, load, reliability, bandwidth, or a mixture of these. In current practice a single path that minimizes the number of hops is used for routing on each origin-destination pair. Research has also considered multipath routing strategies to improve performance by exploiting the available transmission capacity on a set of alternative paths. The first paper on multipath routing was [5] which developed a distributed routing algorithm minimizing total delay in the case of flow dependent latencies but with fixed source rates. For the case of TCP rates that depend on network congestion, the model presented in [2] and further developed in [6], also considered a multipath routing but performed directly by sources. Since distributed routing has scalability advantages, a router-based approach to multipath routing with elastic demands was investigated in [7] and [8].

In addition to increased efficiency, a further argument in favor of multipath routing is stability. Indeed, when route choice is based on metrics that are affected by congestion, such as queuing delay or link latencies, routing and rate control become mutually inter-dependent and equilibrium can be achieved only if both aspects are considered jointly: routing affects the rate control through the induced congestion signals, while rate control induces flows that determine in turn which routes are optimal. If routing is restricted to a single path, congestion effects may lead to route flaps. A remedy for such unstable behavior is to allow flows to split over multiple paths in order to balance their loads. An appropriate tool to capture these interactions between rate control and routing is provided by Wardrop equilibrium. On the other hand, since congestion metrics are subject to estimation errors and random effects it is natural to model routing as a stochastic equilibrium assignment.

The goal of this paper is to propose an analytical framework that provides a theoretical support for cross-layer designs for rate control under a router-based multipath routing. Our model combines rate control modeled by NUM, with a routing strategy based on discrete choice distribution models that lead to a Markovian Traffic Equilibrium (MTE). The latter is a decentralized stochastic version of Wardrop's model. The combination of the NUM and MTE models leads to a system of equations that correspond to the optimality conditions of an equivalent Markovian Network Utility Maximization problem (MNUM), a strictly convex unconstrained program of low dimension where the variables are the link congestion prices. This characterization allows to establish the existence and uniqueness of an equilibrium, and provides a basis for designing decentralized protocols for congestion control with multipath routing.

The paper is structured as follows. Section §II reviews the basic components of our cross-layer approach: we recall the NUM framework for modeling the steady state of TCP protocols and we discuss the concepts of Wardrop equilibrium and Markovian routing. Section §III combines NUM and MTE, introducing the MNUM model for routing and rate control. In §III-A we reduce MNUM to a system of equations involving only the link congestion prices, and then in §III-B we show that these equations admit a variational characterization (D-MNUM) proving the existence of a unique equilibrium state. In §IV we briefly discuss how the model might lead to a cross-layer design of a distributed TCP/IP protocol. We close the paper with comparisons to previous work and some perspectives on future research.

II. Notations and preliminaries

The communication network is modeled by a directed graph $G = (N, A)$, where the nodes $i \in N$ represent origins, destinations and intermediate routers, while the arcs $a \in A$ represent the network links. Each link is characterized by a latency $\lambda_a = s_a(w_a) = \lambda_a^0 + \psi_a(w_a)$ where $\lambda_a^0 \geq 0$ represents a constant propagation delay and $p_a = \psi_a(w_a)$ is the expected queuing delay expressed as a continuous and strictly increasing function $\psi_a : [0, c_a) \to [0, \infty)$ of the traffic w_a on the link, with $\psi_a(0) = 0$ and $c_a \in (0, \infty]$. We also consider a finite set of sources $k \in K$ each one generating a flow rate $x^k \geq 0$ from an origin $s_k \in N$ to a destination $d_k \in N$.

A. Rate control and utility maximization under single path routing

Suppose that each source $k \in K$ routes its flow along a fixed sequence of links $(a_1, \ldots, a_{j_k})$, so that the total traffic on a link a is $w_a = \sum_{k \ni a} x^k$ where the summation is over all the sources $k \in K$ whose route contains that link. Consider the queuing delay $p_a = \psi_a(w_a)$ as a

measure of link congestion and assume that each source $k \in K$ adjusts its rate x^k as a decreasing function of the aggregate queuing delay $q^k = \sum_{a \in k} p_a$ on its route, namely $x^k = f_k(q^k)$, where $f_k : (0, \infty) \to (0, \infty)$ is continuous and strictly decreasing with $f_k(q^k) \to 0$ as $q^k \to \infty$. These equilibrium equations may be written as

$$\begin{aligned} f_k^{-1}(x^k) &= q^k = \sum_{a \in k} p_a = \sum_{a \in k} \psi_a(w_a) \\ &= \sum_{a \in k} \psi_a(\textstyle\sum_{s \ni a} x^s) \end{aligned}$$

which correspond to the optimality conditions for the strictly convex minimization problem

$$\text{(NUM)} \qquad \min_{x \in \mathbb{R}^K} \sum_{a \in A} \Psi_a(\textstyle\sum_{s \ni a} x^s) - \sum_{k \in K} U_k(x^k)$$

where $\Psi_a(\cdot)$ denotes a primitive of $\psi_a(\cdot)$ and $U_k(\cdot)$ a primitive of $f_k^{-1}(\cdot)$. Alternatively, the equations may be stated in terms of the queuing delays as

$$\begin{aligned} \psi_a^{-1}(p_a) &= w_a = \sum_{k \ni a} x^k = \sum_{k \ni a} f_k(q^k) \\ &= \sum_{k \ni a} f_k(\textstyle\sum_{b \in k} p_b) \end{aligned}$$

which correspond to the optimality conditions for the strictly convex dual problem

$$\text{(D-NUM)} \quad \min_{p \in \mathbb{R}^A} \sum_{a \in A} \int_0^{p_a} \psi_a^{-1}(y)\, dy - \sum_{k \in K} F_k(\textstyle\sum_{b \in k} p_b)$$

where $F_k(\cdot)$ is a primitive of $f_k(\cdot)$.

EXAMPLE. Consider the model for TCP Vegas proposed in [9], [4]. For each source k and time t, let W_t^k denote the size of the congestion window and $T_t^k = D^k + q_t^k$ the RTT expressed as the sum of the total propagation delay D^k and the queueing delay q_t^k. A Vegas source estimates D^k as the minimum observed RTT in a time window, and tries to keep the difference between the *expected rate* $\hat{x}_t^k = W_t^k / D^k$ and the *actual rate* $x_t^k = W_t^k / T_t^k$ close to a given value $\alpha^k > 0$. To this end, the congestion window is increased if $\hat{x}_t^k - x_t^k < \alpha^k$, and decreased when $\hat{x}_t^k - x_t^k > \alpha^k$. At equilibrium we must have $\hat{x}^k - x^k = \alpha^k$ which yields the equilibrium rate functions

$$x^k = \frac{\alpha^k D^k}{q^k} \triangleq f_k(q^k).$$

A simple model for the queuing delay can be obtained if we assume that each link has a service capacity $c_a > 0$ and an infinite buffer, so that an M/M/1 model gives the expected queuing delay

$$p_a = \frac{w_a}{c_a(c_a - w_a)} \triangleq \psi_a(w_a).$$

The (NUM) formalism can handle other congestion measures different from the queuing delay and has been used to model the steady state of different TCP protocols, each one characterized by specific maps f_k and ψ_a (see [1], [2], [10], [11], [12], [13]).

B. Routing and traffic equilibrium

We review next some equilibrium models for traffic in congested networks. In this setting the source flows x^k are fixed but may be routed along a set of alternative paths R^k connecting the origin s_k to the destination d_k. The basic modeling principle introduced by Wardrop in [14] is that at equilibrium only paths that are optimal should be used to route flow. We stress that, in contrast with rate control which is based solely on the queuing delays $p_a = \psi_a(w_a)$, route optimality will be measured using the total delays $\lambda_a = \lambda_a^0 + \psi_a(w_a)$.

1) Wardrop equilibrium: Suppose that the flow x^k is split into non-negative path-flows $h_r \geq 0$ so that $x^k = \sum_{r \in R^k} h_r$, and let $w_a = \sum_{r \ni a} h_r$ be the induced total link-flows. Let H denote the set of such *feasible flows* (h, w). An equilibrium [14] is characterized by the fact that only optimal paths are used, namely, for each destination $k \in K$ and each route $r \in R^k$ one has

$$h_r > 0 \Rightarrow c_r = \tau_k \tag{1}$$

where $c_r = \sum_{a \in r} \lambda_a = \sum_{a \in r} s_a(w_a)$ denotes the total delay of the route and $\tau_k = \min_{r \in R^k} c_r$ is the minimum cost faced by source k.

These equilibria were characterized in [15] as the optimal solutions of the convex program

$$\text{(P-W)} \qquad \min_{(h,w) \in H} \sum_{a \in A} \int_0^{w_a} s_a(z)\, dz.$$

Since the feasible set H is compact this problem has optimal solutions, while strict convexity implies that the optimal w is unique. Alternatively, the equilibrium delays $\lambda_a = s_a(w_a)$ are the unique optimal solution of the strictly convex unconstrained dual problem

$$\text{(D-W)} \qquad \min_{\lambda \in \mathbb{R}^A} \sum_{a \in A} \int_{\lambda_a^0}^{\lambda_a} s_a^{-1}(z)\, dz - \sum_{k \in K} x^k \tau_k(\lambda)$$

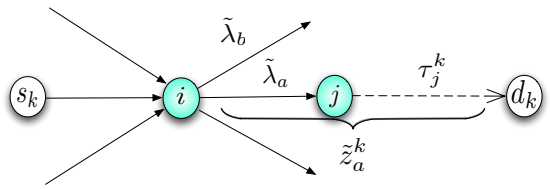

Fig. 1. Variables for dynamic programming equations

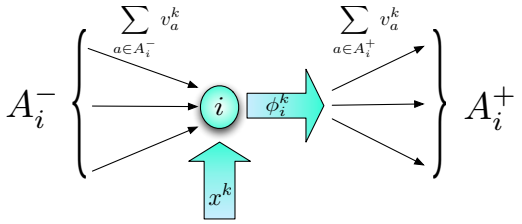

Fig. 2. Flow conservation diagram (here $i = s_k$)

where $\lambda_a^0 = s_a(0)$ and $\tau_k(\lambda) \triangleq \min_{r\in R^k} \sum_{a\in r} \lambda_a$ is the minimum total delay for source $k \in K$.

2) Markovian routing and equilibrium: When link delays are subject to stochastic variability, the route delays $\tilde c_r$ become random variables and the equilibrium conditions (1) are replaced by a stochastic assignment of the form $h_r = x^k\,\mathbb{P}(\tilde c_r \text{ is optimal})$. For instance, if the costs $\tilde c_r$ are i.i.d. Gumbel variables with expected value $c_r = \mathbb{E}(\tilde c_r)$, we get the Logit distribution rule common in the transportation literature

$$h_r = x^k \frac{\exp(-\beta c_r)}{\sum_{p\in R^k} \exp(-\beta c_p)} \qquad (\forall\, r \in R^k)$$

which assigns flow to all the paths, favoring those with smaller expected cost c_r. The parameter β controls how concentrated is the repartition: for $\beta \sim 0$ every path receives an approximately equal share of the flow, while for β large the flow concentrates on paths with minimal cost. Unfortunately, such route-based distribution rules are not amenable to design decentralized scalable protocols.

An alternative is to conceive routing as a stochastic dynamic programming process. Suppose that each packet experiences a random delay $\tilde\lambda_a$ when traversing link a, and let $\tilde\tau_i^k$ be a random variable that represents the total delay from node i to destination d_k. Denote $\lambda_a = \mathbb{E}(\tilde\lambda_a)$ and $\tau_i^k = \mathbb{E}(\tilde\tau_i^k)$ their expected values. If a packet at node i is routed through the link $a \in A_i^+$ we have $\tilde\tau_i^k = \tilde\lambda_a + \tilde\tau_{j_a}^k$, so that a shortest path routing should choose the link with smallest $\tilde\lambda_a + \tilde\tau_{j_a}^k$. Unfortunately, while the link delays $\tilde\lambda_a$ for $a \in A_i^+$ might be observed at node i, this is not the case for the $\tilde\tau_{j_a}^k$'s which depend on future delays that will be experienced when traversing the downstream links. Suppose instead that only the expected values $\tau_{j_a}^k$ are known and available at node i and that each packet from source $k \in K$ observes the $\tilde\lambda_a$'s and is routed through the link $a \in A_i^+$ that minimizes $\tilde z_a^k = \tilde\lambda_a + \tau_{j_a}^k$ to the next node j_a where the process repeats. Thus, denoting $E_a^k \triangleq \{\tilde z_a^k \le \tilde z_b^k \ \forall b \in A_i^+\}$, the packets from source $k \in K$ move across the network according to a Markov chain with transition probabilities

$$P_{ij}^k = \begin{cases} \mathbb{P}(E_a^k) & \text{if } ij = a \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

for $i \neq d_k$, while the destination d_k is an absorbing state. The expected flows correspond to the invariant measures of these Markov chains, leading to a flow distribution rule in which the throughput flow ϕ_i^k from source k that enters node i, splits among the links $a \in A_i^+$ according to (see Figure 2)

$$v_a^k = \phi_i^k\, \mathbb{P}(E_a^k). \qquad (3)$$

The throughputs $\phi^k = (\phi_i^k)_{i\neq d_k}$ can be computed from the stationary equations $\phi^k = \sum_{j=0}^\infty [(\hat P^k)']^j \delta^k x^k$, where $\hat P^k = (P_{ij}^k)_{i,j\neq d_k}$ is the reduced transition matrix on the non-absorbing states, and $\delta_i^k = 1$ for $i = s_k$ and $\delta_i^k = 0$ otherwise. This may also be written as $\phi^k = x^k\delta^k + (\hat P^k)'\phi^k$ which corresponds to the standard flow conservation equations

$$\phi_i^k = x^k\delta_i^k + \textstyle\sum_{a\in A_i^-} v_a^k. \qquad (4)$$

These equations can be restated in compact form using expected utility theory. Namely, let us write $\tilde z_a^k = z_a^k + \epsilon_a^k$ as the sum of its expected value $z_a^k = \lambda_a + \tau_{j_a}^k$ plus a noise ϵ_a^k with $\mathbb{E}(\epsilon_a^k) = 0$, and assume that the distribution of ϵ_a^k does not change with z_a^k (for a discussion of this assumption see §VI). Then, the transition probabilities in (2) can be expressed as $\mathbb{P}(E_a^k) = \frac{\partial\varphi_i^k}{\partial z_a^k}(z^k)$ where φ_i^k denote the expected utility functions

$$\varphi_i^k(z^k) = \begin{cases} \mathbb{E}(\min_{a\in A_i^+}\{z_a^k + \epsilon_a^k\}) & \text{if } i \neq d_k \\ 0 & \text{if } i = d_k \end{cases} \qquad (5)$$

which allow to rewrite the flow equations (3)-(4) as

$$\begin{cases} v_a^k = \phi_i^k \, \frac{\partial \varphi_i^k}{\partial z_a^k}(z^k) & \forall a \in A_i^+ \\ \phi_i^k = x^k \delta_i^k + \sum_{a \in A_i^-} v_a^k & \forall i \neq d_k. \end{cases} \tag{6}$$

On the other hand, assuming that the cost-to-go variables $\{\tilde{\tau}_{j_a}^k : a \in A_i^+\}$ are independent from the local queueing times $\{\tilde{\lambda}_a : a \in A_i^+\}$, we may compute the expected value of $\tilde{\tau}_i^k$ by conditioning on the events E_a^k as

$$\begin{aligned} \tau_i^k = \mathbb{E}(\tilde{\tau}_i^k) &= \sum_{a \in A_i^+} \mathbb{E}(\tilde{\lambda}_a + \tilde{\tau}_{j_a}^k | E_a^k) \mathbb{P}(E_a^k) \\ &= \sum_{a \in A_i^+} \mathbb{E}(\tilde{\lambda}_a + \tau_{j_a}^k | E_a^k) \mathbb{P}(E_a^k) \\ &= \mathbb{E}(\min_{a \in A_i^+} \{\tilde{\lambda}_a + \tau_{j_a}^k\}) \end{aligned}$$

so that

$$\begin{cases} \tau_i^k = \varphi_i^k(z^k) & \forall i \in N \\ z_a^k = \lambda_a + \tau_{j_a}^k & \forall a \in A. \end{cases} \tag{7}$$

Under mild conditions it was proved in [16] that, given the λ_a's, system (6)-(7) has a unique solution (v, ϕ, τ, z). It was also shown that these equations, together with the equilibrium conditions $\lambda_a = s_a(w_a)$ where $w_a = \sum_{k \in K} v_a^k$ represents the total expected link load, have a unique solution $(\lambda, w, v, \phi, \tau, z)$ called a *Markovian Traffic Equilibrium* (MTE). This equilibrium is characterized by a pair of dual optimization problems analog to (P-W) and (D-W). As a matter of fact, the dual problem has exactly the same form

$$\text{(D-MTE)} \qquad \min_{\lambda \in \mathbb{R}^A} \sum_{a \in A} \int_{\lambda_a^0}^{\lambda_a} \psi_a^{-1}(z)\, dz - \sum_{k \in K} x^k \tau_k(\lambda)$$

where $\tau_k(\lambda) \triangleq \tau_{s_k}^k(\lambda)$ with $\tau_i^k(\lambda)$ the solution of (7).

The expected utility maps $\varphi_i^k(\cdot)$ convey all the information required to describe a Markovian routing and may be considered as the primary modeling objects. These maps are determined by the random variables ϵ_a^k which are ultimately tied to the arc random costs $\tilde{\lambda}_a$. However, the class $\mathcal{E}$ of maps that can be expressed in the form (5) admits an analytic characterization (see [16]): they are the $\mathcal{C}^1$ maps $\varphi : \mathbb{R}^n \to \mathbb{R}$ that are concave, componentwise non-decreasing, and which satisfy in addition

(a) $\varphi(x_1 + c, \ldots, x_n + c) = \varphi(x_1, \ldots, x_n) + c$
(b) $\varphi(x) \to x_i$ when $x_j \to \infty$ for all $j \neq i$
(c) for x_i fixed, $\frac{\partial \varphi}{\partial x_i}(x_1, \ldots, x_n)$ is a continuous distribution function on the remaining variables.

We remark also that $\varphi(x) \leq \min\{x_1, \ldots, x_n\}$. In what follows we assume the model is specified directly in terms of a family of maps $\varphi_i^k \in \mathcal{E}$ with $\varphi_{d_k}^k \equiv 0$, though they are not used explicitly by our distributed protocol in §IV.

REMARK. Since packet movements are governed by a Markov chain, cycling may occur and additional conditions are required to ensure that packets reach the destination with probability one. A simple case is when at node i source k considers only the arcs in A_i^+ that lead closer to destination d_k (*e.g.* $\tau_{j_a}^k < \tau_{i_a}^k$), so that the corresponding Markov chain is supported over an acyclic graph (N, A^k). To deal with this case it suffices to redefine

$$\varphi_i^k(z^k) \triangleq \mathbb{E}(\min_{a \in A_i^{k+}} \{z_a^k + \epsilon_a^k\})$$

so that $P_{ij}^k = \frac{\partial \varphi_i^k}{\partial z_a^k}(z^k) = 0$ for all $a = ij \notin A_i^{k+}$.

III. RATE CONTROL WITH MARKOVIAN ROUTING

We proceed to develop a cross-layer model that combines a NUM approach for rate control based on queuing delays, with a Markovian multipath routing based on total delays. Each source $k \in K$ is characterized by an origin s_k, a destination d_k, and a continuous decreasing rate function $f_k : (0, \infty) \to (0, \infty)$ such that $f_k(q^k) \to 0$ when $q^k \to \infty$, while every link $a \in A$ has a continuous increasing latency function $s_a : [0, c_a) \to [\lambda_a^0, \infty)$ with $\lambda_a^0 = s_a(0) \geq 0$. Packets are routed according to a Markovian strategy characterized by a family of maps $\varphi_i^k \in \mathcal{E}$ with $\varphi_{d_k}^k \equiv 0$, and sources adjust their rates as a function $x^k = f_k(q^k)$ of the total queuing delay $q^k = \tau_k(\lambda) - \tau_k^0$, where $\tau_k(\lambda)$ is the end-to-end expected delay as defined in the previous section and $\tau_k^0 = \tau_k(\lambda^0)$ is the free-flow value of this end-to-end delay.

Informally, the source rates x^k induce flows v_a^k and total link loads w_a. These loads determine link expected delays $\lambda_a = s_a(w_a)$ that yield end-to-end delays $\tau_k(\lambda)$ for each source and corresponding queuing delays q^k. At equilibrium, these queuing delays must induce the original rates $x^k = f_k(q^k)$.

Definition 1: A pair (w, x) with $w = (w_a)_{a \in A}$ and $x = (x^k)_{k \in K}$ is called a *Markovian Network Utility Maximization (MNUM)* equilibrium iff $w_a = \sum_{k \in K} v_a^k$ where (v^k, ϕ^k) solve the flow conservation constraints (6) with (τ^k, z^k) satisfying (7), together with the link delay relations $\lambda_a = s_a(w_a)$ and the rate equilibrium conditions $x^k = f_k(q^k)$ where $q^k = \tau_k(\lambda) - \tau_k^0$.

A. Reduced formulation of MNUM

In order to establish the existence and uniqueness of equilibria we begin by reducing MNUM to an equivalent set of equations that involves only the variables λ. To this end we need to extend the results in [16]. We omitted in this publication full proofs, but we refer to [17] for them. Consider a fixed non-negative link delay vector $(\lambda_a)_{a\in A}$. We first show that (7) uniquely defines z^k and τ^k as implicit functions of λ. This system can be equivalently stated solely in terms of the variables τ^k as

$$\tau_i^k = \varphi_i^k\left((\lambda_a + \tau_{j_a}^k)_{a\in A}\right) \tag{8}$$

so it suffices to prove that the latter uniquely defines τ^k as a function of λ.

Proposition 1: Let $k \in K$ and denote $\bar\tau_i^k$ the cost of a shortest path from i to destination d_k with link costs λ_a. Suppose also that $\hat\tau^k \in \mathbb{R}^N$ is such that

$$\hat\tau_i^k \le \varphi_i^k((\lambda_a + \hat\tau_{j_a}^k)_{a\in A}) \qquad (\forall\, i \in N). \tag{9}$$

Then $\hat\tau^k \le \bar\tau^k$ and moreover, starting from $\tau^{k,0} = \bar\tau^k$, the iterates computed by

$$\tau_i^{k,n+1} = \varphi_i^k((\lambda_a + \tau_{j_a}^{k,n})_{a\in A}) \tag{10}$$

are non-increasing and converge for $n \to \infty$ to a solution τ^k of (8) with $\tau_i^k \in [\hat\tau_i^k, \bar\tau_i^k]$.

REMARK. Observe that if we know φ_i^k, the previous result gives a procedure to solve (8): compute the shortest path delays $\bar\tau^k$ and then iterate (10). Alternatively one may start from $\tau^{k,0} = \hat\tau^k$ in which case the iterates increase and are bounded from above by $\bar\tau^k$, hence these iterates also converge to a solution of (8).

We present next an extension of a result from [16] which is the basis to prove uniqueness. In what follows we denote $\mathcal{P}$ the *open* convex domain of all $\lambda \in \mathbb{R}^A$ for which there exists $\hat\tau^k \in \mathbb{R}^K$ satisfying

$$\hat\tau_i^k < \varphi_i^k((\lambda_a + \hat\tau_{j_a}^k)_{a\in A}) \text{ for all } i \ne d_k. \tag{11}$$

Note that if $\lambda \in \mathcal{P}$ then $\lambda' \in \mathcal{P}$ for all $\lambda' \ge \lambda$.

Lemma 1: Let $\lambda \in \mathcal{P}$ and suppose that (τ^k, z^k) solves (7). Let $\hat Q^k(z^k)$ be the matrix with entries $\hat Q_{ia}^k(z^k) = \frac{\partial\varphi_i^k}{\partial z_a^k}(z^k)$ for $i \ne d_k$ and $a \in A$. Then

(a) For each $i \ne d_k$ there exists $j \in N$ such that $P_{ij}^k > 0$ and $\hat\tau_j^k - \tau_j^k > \hat\tau_i^k - \tau_i^k$.
(b) The matrix $[I - \hat P^k(z^k)]$ is invertible.
(c) Equation (6) has a unique solution given by $v^k = \hat Q^k(z^k)'\phi^k \ge 0$ with $\phi^k = [I - \hat P^k(z^k)']^{-1}\delta^k x^k \ge 0$.

The next result is the key to reduce the MNUM equations to a system in the variables λ.

Proposition 2: If $\lambda \in \mathcal{P}$ then, for each source $k \in K$, the system (7) has a unique solution $z^k = z^k(\lambda) > 0$ and $\tau^k = \tau^k(\lambda) > 0$. Moreover, the functions $\lambda \to \tau_i^k(\lambda)$ and $\lambda \to z_a^k(\lambda)$ are concave, smooth and component-wise non-decreasing.

The implicit maps $\tau^k(\lambda)$ and $z^k(\lambda)$ defined by (7), allow to restate the MNUM equations solely in terms of the link delay vector λ. Indeed, let $q^k(\lambda) = \tau_{s_k}^k(\lambda) - \tau_k^0$ and define $\tilde x^k(\lambda) = f_k(q^k(\lambda))$. According to Lemma 1(c) the equations (6) have unique solutions $v^k = v^k(\lambda)$ and $\phi^k = \phi^k(\lambda)$. Denoting $\tilde w_a(\lambda) = \sum_{k\in K} v_a^k(\lambda)$, the MNUM equations are equivalent to the reduced system of equations

$$\text{(R-MNUM)} \qquad \lambda_a = s_a(\tilde w_a(\lambda)) \quad \forall a \in A.$$

B. Variational characterization

We show that the reduced system (R-MNUM) corresponds to the optimality conditions of an optimization problem which is a combination of the variational characterizations (D-NUM) and (D-MTE).

Theorem 1: Assume that $\lambda^0 \in \mathcal{P}$. Then (x^*, w^*) is an MNUM equilibium iff $x^* = \tilde x(\lambda^*)$ and $w^* = \tilde w(\lambda^*)$ where λ^* is the unique optimal solution of the strictly convex program

(D-MNUM)

$$\min_{\lambda\in\mathcal{P}}\ \Phi(\lambda) \ \triangleq\ \sum_{a\in A}\int_{\lambda_a^0}^{\lambda_a} s_a^{-1}(z)\,dz - \sum_{k\in K} F_k(q^k(\lambda))$$

with $F_k(\cdot)$ a primitive of $f_k(\cdot)$.

This characterization allows us to prove the existence and uniqueness of an MNUM equilibrium.

Theorem 2: Problem (D-MNUM) is strictly convex and coercive, hence it has a unique optimal solution and therefore there exists a unique MNUM equilibrium.

IV. A DISTRIBUTED ALGORITHM FOR MNUM

This section describes how the MNUM framework can lead to a distributed protocol for rate congestion control under Markovian multipath routing. This protocol can be interpreted as a distributed algorithm that solves

the variational problem (D-MNUM). The algorithm is based on a Markovian routing process for packets, with a slow update of the end-to-end expected delays τ_i^k's. This process is combined with a fast TCP adaptation of user's rates by estimating the end-to-end queuing delays to reach the equilibrium rates $\tilde{x}^k(\lambda)$. A more detailed description and analysis of the distributed protocol will be the subject of a forthcoming paper [18].

A. Packet routing based on local queues

We adapt the ideas of §II-B2 in order to define a routing policy based on local information. To do this, router i needs to observe the values $\tilde{z}_a^k = \tilde{\lambda}_a + \tau_{j_a}^k$ of each outgoing link $a \in A_i^+$ by adding the link propagation delay λ_a^0, plus the current queueing delay $\tilde{p}_a$ of the link, plus the estimate of the expected delay $\tau_{j_a}^k$ informed by the next hop.

Routers must periodically provide their neighboring routers with an estimate of the total expected delays τ_i^k. These estimates may be updated on a slow time-scale by averaging the observed delays $\min_{a\in A_i^+}\{\tilde{\lambda}_a + \tau_{j_a}^k\}$ over all packets routed on a fixed time window or a fixed number of packets. The observed average $\check{\tau}_i^k$ is used to update the estimate of the expected delay as

$$\tau_i^k \leftarrow (1-\alpha)\tau_i^k + \alpha\check{\tau}_i^k.$$

B. TCP protocol

A TCP protocol for source rate control requires a consistent feedback congestion signal. Basically, we require a mechanism by which every source k can estimate $q^k(\lambda)$. Here we rely on the two time-scales assumption: sources control their rates at a much faster pace than routers, so they see link delays as constant. For fixed expected delays λ, the expected forward time coincides with $\tau_{s_k}^k(\lambda)$, thus using standard protocols for estimating the forward time, users can get an unbiased estimation T_t^k of this total expected delay.

Finally, we need to compute the free-flow total expected delay τ_k^0. Since the maps $\varphi_i^k(\cdot)$ are not known explicitly, it is not possible to solve the equations (7) directly. We propose to estimate τ_k^0 by the minimum observed time for every packet, as in the single-path implementation of Vegas. This provides a biased lower bound estimate for τ_k^0.

The free-flow times τ_k^0 together with the unbiased estimators T_t^k of $\tau_{s_k}^k$ for every packet t arriving to destination, can be used to adjust the rates by a stochastic approximation algorithm of the form

$$x_{t+1}^k \leftarrow (1-\delta)x_t^k + \delta f_k(Q_t^k) \tag{12}$$

where $Q_t^k = T_t^k - \tau_k^0$. If we let sources adapt long enough so that $Q_t^k \sim q^k$ and $x_t^k \sim f_k(q^k)$, we can then proceed to update the router estimates of the end-to-end delays τ_i^k. Further considerations on how multipath routing affects buffering at the destination and other issues will be addressed in [18].

V. Comparison with previous work

As mentioned in the Introduction there have been several proposals to develop multipath routing protocols, such as [5], [2] and [6]. However, our work is more closely connected to [7], [8] which describes an approach for rate congestion control under a distributed multipath routing protocol based on routing proportions $(\alpha_a^k)_{a\in A_i^+}$ that control the routing of packets from source k at node i. These proportions are dynamically adjusted so that eventually the routing concentrates over the links $a \in A_i^+$ that belong to shortest expected paths.

From a conceptual point of view there are two main differences between this approach and ours. Firstly, the proportion-based approach makes routing decisions based on expected values, while our splitting of traffic evolves stochastically as it uses the current state of local queues to optimize the routing. A second difference is that [7], [8] optimize routing in terms of queuing delay only, while our routing optimization considers the total delay including queuing plus propagation delay. This choice makes sense if one cares about the total time that packets take to be transmitted, and not just the time spent in queues.

Further differences arise at the implementation level. The protocol in [7], [8] requires three time-scales: a fast TCP rate adaptation, a medium time-scale for route price updates, and a slow update for flow-splitting proportions. This time-scale separation is required to justify the convergence of the protocol. In our case we only need two time-scales: a slow one for estimating the total delays to destinations and a fast one for TCP rate control and routing.

Finally, in terms of information and communication overhead both implementations are similar, with the difference that in our case we require a fast interaction with local queues in order to choose the best route.

VI. Conclusions and future work

We proposed a new cross-layering model for TCP/IP control under multipath routing. The motivation for our routing mechanism comes from using local information about queueing delays as well as the expected delays from the next hops to the destination, in order to exploit the available capacity by sending packets through several alternative routes. To achieve this purpose, we considered a Markovian routing combined with a Vegas-like TCP protocol for rate control. The routing process was characterized by studying the expected dynamic programming equations which lead to a Markovian Traffic Equilibrium, together with a standard Network Utility Maximization model for the TCP steady state. This led to a variational characterization of the equilibrium that allowed to prove its existence and uniqueness, and which inspired a distributed protocol for attaining the equilibrium.

There are several unsolved issues. Firstly, further research is required to provide a theoretical support for the convergence of these protocols. A detailed analysis should study the relation between the packet-level dynamics and our flow-level model. Our equilibrium model relies on this assumption: namely, we base our updates on aggregated flow information as well as in the two time-scales convergence of equilibrium flows. A complete analysis should explain to which extent the flow model captures the packet level dynamics, and how fast the equilibrium flows are attained by sources. Interesting recent results along this line can be found in [19], [20].

Another interesting question is related to the model of randomness assumed. We considered an additive structure $\tilde{z}_a^k = z_a^k + \epsilon_a^k$ which presumes the same variability of delays regardless of the average flow levels observed. A more realistic model should consider higher variability for higher expected delays, based either on a detailed analysis of the distribution of waiting times at queues, or at least using a simplified multiplicative randomness model of the form $\tilde{z}_a^k = z_a^k(1 + \epsilon_a^k)$. These sophisticated models would justify the use of the minimum observed delay for τ_k^0, since when there is no congestion at all, we can safely assume zero variability of travel times.

A final line of research has to do with simulating this protocol in a realistic environment. A fair comparison with single-path routing requires the presence of uncertainty and delays in information transmission. Simulation may provide an idea on the effective increase in performance that one might expect from a Markovian multipath routing.

References

[1] S. Low, F. Paganini, and J. Doyle, "Internet congestion control," *IEEE Control Systems Magazine*, vol. 22, no. 1, pp. 28–43, Feb. 2002.

[2] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, pp. 237–252, Jan 1998.

[3] H. Yaïche, R. Mazumdar, and C. Rosenberg, "A game theoretic framework for rate allocation and charging of available bit rate (abr) connections in atm networks," in *Broadband Communications*, 1998, pp. 222–233.

[4] S. Low, "A duality model of tcp and queue management algorithms," *IEEE/ACM Transactions on Networking*, pp. 525–536, Jan 2003.

[5] R. Gallager, "A minimum delay routing algorithm using distributed computation," *IEEE Transactions on Communications*, pp. 73–85, Jul 1977.

[6] X. Lin and N. Shroff, "Utility maximization for communication networks with multipath routing," *IEEE Transactions on Automatic Control*, pp. 766–781, Jan 2006.

[7] F. Paganini, "Congestion control with adaptive multipath routing based on optimization," in *Information Sciences and Systems*, Jan 2006, pp. 333–338.

[8] E. Mallada and F. Paganini, "Optimal congestion control with multipath routing using tcp-fast and a variant of rip," *Lecture Notes in Computer Science*, pp. 205–214, Jan 2007.

[9] S. Low, L. Peterson, and L. Wang, "Understanding vegas: a duality model," *Journal of the ACM*, vol. 49, pp. 207–235, March 2002.

[10] R. Gibbens and F. Kelly, "Resource pricing and the evolution of congestion control," *Automatica*, pp. 1969–1985, Jan 1999.

[11] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, "Modeling tcp reno performance: a simple model and its empirical validation," *IEEE/ACM Transactions on Networking*, vol. 8, pp. 133–145, 2000.

[12] V. Dumas, F. Guillemin, and P. Robert, "A markovian analysis of additive-increase multiplicative-decrease (aimd) algorithms," in *Advances in Applied Probability*, 2002, pp. 85–111.

[13] S. Kunniyur and R. Srikant, "End-to-end congestion control schemes: utility functions, random losses and ecn marks," *IEEE/ACM Transactions on Networking*, vol. 11, pp. 689–702, October 2003.

[14] J. G. Wardrop, "Some theoretical aspects of road traffic research," *Proceedings of the Institute of Civil Engineers, Part II*, pp. 325–378, 1952.

[15] M. Beckman, C. McGuire, and C. Winsten, *Studies in Economics of Transportation*. Yale University Press, January 1956.

[16] J. Baillon and R. Cominetti, "Markovian traffic equilibrium," *Mathematical Programming*, vol. 111, pp. 33–56, Jan 2008.

[17] R. Cominetti and C. Guzman, "Network Congestion Control with Markovian Multipath Routing," *ArXiv e-prints*, Jul. 2011.

[18] C. Guzmán, "Implementation of a distributed protocol for network congestion control with markovian multipath routing," *Forthcoming*, p. na, 2012.

[19] N. S. Walton, "Proportional fairness and its relationship with multi-class queueing networks," *Annals of Applied Probability*, pp. 2301–2333, 2009.

[20] F. Kelly, L. Massoulié, and N. Walton, "Resource pooling in congested networks: proportional fairness and product form," *Queueing Systems*, vol. 63, no. 1-4, pp. 165–194, 2009.