Congestion control for high performance, stability and fairness in general networks

Fernando Paganini Zhikui Wang John C. Doyle Steven H. Low

***Abstract*— This paper is aimed at designing a congestion control system that scales gracefully with network capacity, providing high utilization, low queueing delay, dynamic stability, and fairness among users. The focus is on developing decentralized control laws at end-systems and routers at the level of fluid-flow models, that can provably satisfy such properties in arbitrary networks, and subsequently approximate these features through practical packet-level implementations.**

Two families of control laws are developed. The first "dual" control law is able to achieve the first three objectives for arbitrary networks and delays, but is forced to constrain the resource allocation policy. We subsequently develop a "primal-dual" law that overcomes this limitation and allows sources to match their steady-state preferences at a slower time-scale, provided a bound on round-trip-times is known.

We develop two packet-level implementations of this protocol, using (i) ECN marking, and (ii) queueing delay, as means of communicating the congestion measure from links to sources. We demonstrate using ns-2 simulations the stability of the protocol and its equilibrium features in terms of utilization, queueing and fairness, under a variety of scaling parameters.

I. Introduction

The congestion control mechanisms in the Internet consist of the congestion window algorithms of TCP [16], running at end-systems, and active queue management (AQM) algorithms (e.g. [11]) at routers, seeking to obtain high network utilization, small amounts of queueing delay, and some degree of fairness among users. These implementations are the result of an evolutionary cycle involving heuristics, small-scale simulation and experimentation, and deployment; given that this process occurred at the time of an explosive growth of the network, the achieved performance of these systems must be considered a resounding success. However, there are reasons to wonder whether this evolutionary path may be reaching its limits: deficiencies of the current loss-based protocol in wireless environments; difficulties in providing quality of service guarantees (delay, resource allocation); and the growing evidence (see e.g. [26], [12], [19]) that the additive-increase-multiplicative- decrease (AIMD) structure in TCP does not scale well to high capacity networks.

The role of mathematical models in this design process has been modest: perhaps the most cited early reference is [17], which gives a quasi-static analysis of fairness properties of a highly idealized version of AIMD. Mathematical explanations of the *dynamics* of TCP/AQM have only recently been pursued (e.g., [10], [9], [26], [15]), and they typically only have predictive value in very simple scenarios such as single bottlenecks with homogeneous delays. Indeed, the complexity of the nonlinear delay-differential equations that arise should quickly convince anyone of the intractability of making predictions at the global scale. Superficially, this seems to confirm that mathematical models have limited value so the empirical route was the only alternative. However, to some degree this is a self-fulfilling prophecy: as in other complex systems, seeking mathematical verification *a posteriori* to a heuristic design is rarely tractable; but sometimes a rigorous foundation can be attained if one "designs for provability". Strikingly, it has recently become clear that such foundation is available for the congestion control problem, within the *same* design principles that have guided the Internet (end-to-end control, no per-flow state in the network, see [5]), and only requiring minor modifications to the details of the algorithms. This formulation originates in the work of [20], [13], [21], and is based on fluid-flow models and the explicit consideration of a *congestion measure* fed back to sources from congested links. Interpreting such signals as *prices* has allowed for economic interpretations that make explicit the equilibrium resource allocation policy specified by the control algorithms, in terms of a suitable optimization problem [20], [25]. In terms of dynamics, these models also reveal a special structure that can be exploited for control design, as pursued recently in [18], [28], [29], [32], [22]. The present paper gives a comprehensive treatment of one such approach; preliminary versions of this work are reported in the conference papers [29], [30].

We pose the objective of finding a protocol that can be implemented in a decentralized way by sources and routers, and controls the system to a stable equilibrium point which satisfies some basic requirements: high utilization of network resources, small queues, and a degree of control over resource allocation. All of these are required to be *scalable*, i.e. hold for an arbitrary network, with possibly high capacity and delay. This fact, and the decentralized information structure, significantly narrow down the search for a control law. In Section III we present a first candidate solution that is able to achieve the first two equilibrium objectives, and stability, in a completely scalable way, but constrains the resource allocation policy. In Section IV we extend the theory to include dynamics at TCP sources, preserving the earlier features at fast time-scales (high frequencies) but permitting sources also to match their steady-state (low frequency) preferences; the only limitation to scalability is that a bound on round-trip-

F. Paganini and Z. Wang are with University of California, Los Angeles; emails: {paganini,zkwang}@ee.ucla.edu. S. Low and J. Doyle are with the California Institute of Technology; emails: {slow@its,doyle@cds}.caltech.edu. Research supported by NSF Award ECS-9875056, the David and Lucille Packard Foundation, and the DARPA-ITO NMS program.

times is assumed to be known. Using time-scale separation in these problems was originally considered in [22], and further pursued in [23]; our results are in a sense dual to those in the latter reference.

In the final sections of the paper we describe how to go beyond fluid-flow models and pursue a packet-level protocol with these features, within the constraints of mechanisms currently available in the Internet. Two strategies are pursued: one, described in Section V, is based on the Explicit Congestion Notification (ECN) bit available in the packet header to code the congestion information between links and sources; this version has the advantage of allowing operation with essentially zero delay, at the cost of some added complexity in network routers. We present some ns-2 simulation tests to demonstrate the performance, in highly stressed congestion scenarios and high capacity links. The second implementation, described in Section VI, is based on queueing delay as a congestion measure, similar to what is done in TCP Vegas [3]. This allows some degradation of performance in terms of queueing delay and fairness, but has the advantage of requiring no explicit participation from routers.

Conclusions are given in Section VII, and some proofs are given in the Appendix.

II. Preliminaries

A. Fluid-flow models for congestion control

We are concerned with a system of communication links, indexed by l, shared by a set of source-destination pairs, indexed by i. The routing matrix R is defined by

$$R_{li} = \begin{cases} 1 & \text{if source } i \text{ uses link } l \\ 0 & \text{otherwise} \end{cases},$$

and assumed fixed. The theory will be based on a fluid-flow abstraction of the TCP/AQM congestion control problem. Each source i has an associated transmission rate $x_i(t)$; the set of transmission rates determines the aggregate flow $y_l(t)$ at each link, by the equation

$$y_l(t) = \sum_i R_{li} x_i(t - \tau_{li}^f), \tag{1}$$

in which the forward delays τ_{li}^f between sources and links are accounted for. Each link has a capacity c_l in packets per second.

In practice, routing varies as sources arrive or leave the network, or based on routing changes, but we assume this happens at a slower time-scale than our analysis. We remark that we are modeling only persistent sources which can be controlled. From the point of view of these "elephants", what matters is settling on a set of rates which fully utilizes the network bandwidth and distributes it appropriately among them. The network is also shared by short "mice", which don't last long enough to be controlled, but are affected by the elephant dynamics, mainly through the queuing delay they experience. We will not model them explicitly here (they could be treated as noise in link rates), but will bear their objectives in mind for design.

The feedback mechanism is modeled as follows [20], [25]: each link has an associated congestion measure or *price* $p_l(t)$, and sources are assumed to have access to the *aggregate* price of all links in their route,

$$q_i(t) = \sum_l R_{li} p_l(t - \tau_{li}^b). \tag{2}$$

Here again we allow for backward delays τ_{li}^b in the feedback path from links to sources. As discussed in [25], [27], this feedback model includes, to a good approximation, the mechanism present in existing protocols, with a different interpretation for price in different protocols (e.g. loss probability in TCP Reno, queueing delay in TCP Vegas). The vectors x, y, p, q collect the above quantities across sources and links.

The total RTT for the source thus satisfies

$$\tau_i = \tau_{li}^f + \tau_{li}^b \tag{3}$$

for every link in the source's path. These delays contain a fixed component of propagation and packet processing, but could also include *queueing delays*, which vary dynamically in time. When necessary, we will denote by d_i the fixed portion of the round-trip-time.

In this framework, a congestion control system is specified by choosing (i) how the links fix their prices based on link utilization; (ii) how the sources fix their rates based on their aggregate price. These operations will determine both the equilibrium and dynamic characteristics of the overall system.

B. Equilibrium objectives and utility-based interpretation

We first specify the design objectives for the equilibrium point to be achieved by our system:

1) Network utilization. Link equilibrium rates y_{0l} should of course not exceed the capacity c_l, but also should attempt to track it.
2) Equilibrium queues should be empty (or small) to avoid queueing delays.
3) Resource allocation. We will assume sources have a demand curve
$$x_{0i} = f_i(q_{0i}) \tag{4}$$
that specifies their desired equilibrium rate as a decreasing function of price. This is equivalent to assigning them a concave *utility function* $U_i(x_i)$, in the language of [20], and postulating that sources choose their equilibrium rate from their local maximization of "profit",
$$\max_{x_{0i}} [U_i(x_{0i}) - q_{0i} x_{0i}].$$
This gives the relationship (4) with $f_i = (U_i')^{-1}$. We would like the control system to reach an equilibrium that accommodates these demands. The choice of utility function provides a "soft" way of imposing fairness (weaker than, e.g."max-min" fairness [2]), or alternatively service differentiation; this market-based approach is consistent with the end-to-end philosophy [21].

Although the control design will be described in detail in the following sections, it is useful to introduce right away the type of algorithm to be used at the links. Consider the mechanism

$$\dot{p}_l = \begin{cases} \gamma_l(y_l - c_{0l}), & \text{if } p_l > 0 \text{ or } y_l > c_{0l}; \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

to set prices at each link. Here c_{0l} is a target capacity, and γ_l a positive constant to be chosen.

Clearly, at any equilibrium point we will have $y_{0l} \leq c_{0l}$, and the price will be nonzero only at bottleneck links where $y_{0l} = c_{0l}$. If we choose $c_{0l} = c_l$, the capacity would be matched at certain bottlenecks, and every source would see a bottleneck (assuming its demand function is able to fill it). So the above algorithm, if it reaches equilibrium, would satisfy our utilization objective.

This kind of link algorithm was studied in [25] and related to the optimization of total utility subject to network capacity constraints:

$$\max_{x \geq 0} \sum_i U_i(x_i), \quad \text{subject to} \quad Rx \leq c\ . \tag{6}$$

An equilibrium point of (5) together with a source algorithm that satisfies (4) is a solution to the above convex program; furthermore, the equilibrium prices are the Lagrange multipliers for the corresponding dual.

The main drawback of choosing $c_{0l} = c_l$ is that it leads to nonzero equilibrium queues. Indeed, a simple fluid-flow model for a backlog or queue b_l at a link is the equation

$$\dot{b}_l = \begin{cases} y_l - c_l, & \text{if } b_l > 0 \text{ or } y_l > c_l; \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Namely, the queue or backlog b_l in packets integrates the excess rate over capacity, and is always non-negative[1]. Comparing to (5), we see that prices would be proportional to queues and thus bottleneck links would have a backlog. This leaves two options: one can either work with other design parameters to make sure this backlog is small, or instead, make c_{0l} a "virtual" capacity slightly below c_l (an idea from [14]). In this way equilibrium queues can be empty with bottleneck links at essentially full utilization[2].

C. Dynamic objectives and a linearized model

Equilibrium considerations are meaningful if the system can operate at or near this point; for this reason we pose as a basic requirement the stability of the equilibrium point. Ideally, we would seek global stability from any initial condition, but at the very least we should require local stability from a neighborhood.

This objective is sometimes debated, since instability in the form of oscillations could perhaps be tolerated in the network context, and might be a small price to pay for an aggressive control. We emphasize, however, that going beyond the limit of stability one sacrifices any predictability of system behavior, and any evidence that oscillations are benign would inevitably be anecdotal. Examples where they are quite severe can be found in [9], [26]. Instead, making a system stable but close to the stability boundary, as will be pursued below, provides the maximum speed of response compatible with a predictable steady state operation.

In this paper we will only pursue *local* stability results, based on small perturbations $x = x_0 + \delta x$, $y = y_0 + \delta y$, $p = p_0 + \delta p$, $q = q_0 + \delta q$ around equilibrium, and studied via linearization. We assume that links are running the control law (5), and for most of the theory we will assume $c_{0l} < c_l$. This has the following implications:

- Around equilibrium, there is no queueing delay. This means the delays $\tau^f_{i,l}, \tau^b_{i,l}$ only have their fixed component, therefore (1) and (2) are linear time invariant relationships, amenable to study via the Laplace transform.
- No links are "truly" saturated. This means an increase δx_i in a certain source's rate will be seen by *all* the bottlenecks in its path.
- Non-bottleneck links have zero price, a fact not affected by a small perturbation. Thus δp_l will only be nonzero for bottlenecks, and we can reduce the analysis to such links.

With these considerations, we can linearize (1-2) and express the result in the Laplace domain, as follows:

$$\delta\bar{y}(s) = \bar{R}_f(s)\delta x(s), \tag{8}$$

$$\delta q(s) = \bar{R}_b(s)^T \delta\bar{p}(s). \tag{9}$$

Here we use the notation $\delta\bar{p}, \delta\bar{y}$ to indicate the reduced vectors obtained by eliminating non-bottleneck links. Also, the matrices $\bar{R}_f(s)$ and $\bar{R}_b(s)$ are obtained by eliminating non-bottleneck rows from R, and also replacing the "1" elements respectively by the delay terms $e^{-\tau^f_{i,l}s}$, $e^{-\tau^b_{i,l}s}$. The superscript ^{T} denotes transpose.

We will assume that the matrix $\bar{R} = \bar{R}_f(0) = \bar{R}_b(0)$ is of full row rank. This means that there are no algebraic constraints between bottleneck link flows, ruling out, for instance, the situation where all flows through one link also go through another. Typically, however, in that situation only one of the links would be a bottleneck; so our assumption is quite generic.

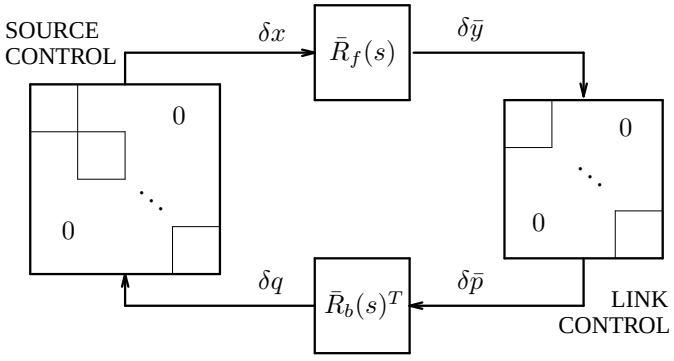

Fig. 1. General congestion control structure.

With these conventions, we can represent the linearized feedback system in the block diagram of Fig. 1. Here, the network portion is represented by the matrices $\bar{R}_f(s)$, $\bar{R}_b(s)^T$, which encode the routing and delay information. These are

[1] Other models [20], inspired in steady-state stochastic queueing theory, treat queues as static functions of the rate which grow as it approaches capacity. The above integrator model seems more appropriate for dynamic studies when queues spend a significant proportion of time in the non-empty state.

[2] Another approach used in [1] is to add another "integrator" to the price dynamics; this, however, poses limitations on scalable stability so it will not be pursued here.

given and fixed, but not known to the sources and links. Moreover, the latter operate in a decentralized manner based only on local information, as represented in the figure by the block-diagonal structure. These tight information constraints make this a challenging control design problem.

Following [20], it has become customary to denote by "primal" those control laws that contain dynamics at sources, but static functions at links, and "dual" those laws where the opposite holds. In this vein, we will name "primal-dual" the case where both control laws are dynamic.

III. A "DUAL" CONTROL WITH SCALABLE STABILITY

We first describe a control strategy that is based on (5) at the links, plus a *static* control law

$$x_i = f_i(q_i) \tag{10}$$

at the sources; this means the source follows instantaneously its demand function (4). As such, this is a dual control law of the type studied in [25]. Our aim here is to find a control that would scale itself to achieve local stability for arbitrary networks and arbitrary values of the RTT. This requires a careful choice of the parameter γ_l in (5), and of the function f_i, as is now described.

A. Linearized design and stability theorem

Consider the linearization of (10) around a certain equilibrium point,

$$\delta x_i = -\kappa_i \delta q_i; \tag{11}$$

the negative sign is used since the demand function is decreasing. Also consider the linearization of the link law (5) around a nonzero equilibrium price, in the Laplace domain:

$$\delta \bar{p}_l = \frac{\gamma_l}{s} \delta \bar{y}_l. \tag{12}$$

We will employ matrix notation to describe the overall feedback system; throughout, the notation $\mathrm{diag}(\cdot)$ denotes a diagonal matrix with the corresponding entries on the diagonal.

We first introduce $\mathcal{C} = \mathrm{diag}(\gamma_l)$, $\mathcal{K} = \mathrm{diag}(\kappa_i)$ and express (11-12) as

$$\delta x = -\mathcal{K} \delta q, \quad \delta \bar{p} = \mathcal{C} \frac{I}{s} \delta \bar{y}.$$

Here the matrix of integrators $\frac{I}{s}$ has the dimension of the number of bottleneck links. Combining these laws with the network equations (8-9) as in Fig. 1, we can represent the feedback loop as the standard unity feedback configuration of Fig. 2, with loop transfer function matrix

$$L(s) = \bar{R}_f(s) \mathcal{K} \bar{R}_b^T(s) \mathcal{C} \frac{I}{s}. \tag{13}$$

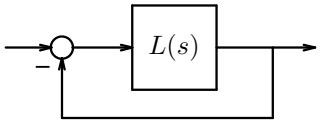

Fig. 2. Overall feedback loop

The negative feedback sign has been pulled out of (11) as is the standard convention; the external input is not relevant to the stability analysis, but it could represent here noise traffic from uncontrolled sources added at the links. The design question is how to choose the gains γ_l and κ_i so that the feedback loop remains stable for arbitrary network topologies, parameters, and delays.

To guide our search, focus first on a single link and source. Here the feedback loop is scalar, with loop transfer function

$$L(s) = \kappa\gamma \frac{e^{-\tau s}}{s},$$

and the stability of the closed loop can be studied with methods of classical control (e.g., [24]), to determine whether the equation $1 + L(s) = 0$ has roots in $Re(s) \geq 0$, which indicates instability. Among the various methods, the *Nyquist criterion* states that our system will be stable as long as the curve $L(j\omega)$ (called the Nyquist plot) does not go through, or encircle, the point -1. The key advantage of this method is that allows one to infer right half-plane information by only looking at the imaginary axis $s = j\omega$; this is particularly useful for the case of time-delay systems that include complex exponentials.

It is not difficult to see that this loop would be unstable for large τ, unless the gain $\kappa\gamma$ compensates for it. Fortunately, sources can measure their RTT so we can set $\kappa = \frac{\alpha}{\tau}$, which gives a loop transfer function

$$L(s) = \alpha\gamma \frac{e^{-\tau s}}{\tau s}. \tag{14}$$

We call the above transfer function, in which the variable s is always multiplied by τ, *scale-invariant*: this means that Nyquist plots for all values of τ would fall on a single curve Γ, depicted in Fig. 3 for $\alpha\gamma = 1$. In the time domain, closed loop responses for different τ's would be the same except for time-scale.

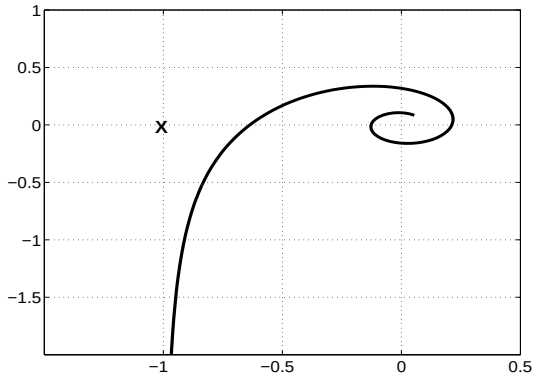

Fig. 3. Nyquist plot Γ of $e^{-j\theta}/j\theta$.

Γ first touches the negative real axis at the point $-2/\pi$, to the right of the critical point -1, so the Nyquist criterion implies that our loop achieves scalable stability for all τ provided that the gain $\alpha\gamma < \pi/2$.

For a single link/source, the preceding gain condition could be imposed a priori. Suppose now that we have N identical sources sharing a bottleneck link. It is not difficult to see that the effective loop gain is scaled up by N; this must be

compensated for if we want stability, but in these networks neither sources nor links know what N is: how can they do the right "gain-scheduling"?

The key idea in our solution is to exploit the conservation law $c_{0l} = \sum_i R_{li} x_{0i}$ implicit in the network equilibrium point, by choosing $\gamma_l = \dfrac{1}{c_{0l}}$ at each link, and making κ_i proportional to $\frac{x_{0i}}{\tau_i}$ at each source. In the case of a single link, but now many sources with heterogeneous delays, this gives a loop transfer function (still scalar, seen from the link as in Fig. 2) of

$$L(j\omega) = \sum_i \frac{x_{0i}}{c_{0l}} \frac{e^{-j\tau_i\omega}}{\tau_i\omega},$$

which gives, at any frequency a *convex combination* of points in Γ. It follows from Fig. 3 that this convex combination will remain on the correct side of the critical point and thus the loop is stable.

Will this strategy work if there are multiple bottleneck links contributing to the feedback? Intuitively, there could be an analogous increase in gain that must be compensated for. Therefore we introduce a gain $\frac{1}{M_i}$ at each source, M_i being a bound on the number of bottleneck links in the source's path, which we assume is available to sources (see Section V). This leads to a local source controller

$$\delta x_i = -\kappa_i \delta q_i = -\frac{\alpha_i x_{0i}}{M_i \tau_i} \delta q_i, \tag{15}$$

where $\alpha_i < \pi/2$ is a constant gain parameter. We have the following result.

Theorem 1: Suppose the matrix $\bar{R} := \bar{R}_f(0) = \bar{R}_b(0)$ is of full row rank, and that $\alpha_i < \frac{\pi}{2}$. Then the system with source controllers (15) and link controllers (12) is linearly stable for arbitrary delays and link capacities.

The proof of this theorem is based on multivariable extension of the above Nyquist argument, and is given in the Appendix. Historically, it was the seminal paper [18] that first introduced linear multivariable stability analysis for the study of delays in the control laws of [20]; motivated by this, parallel independent work in [31], [28], [29], [32] brought in more control theory to further develop these results and to seek *scalable* control laws. The above statement is taken from [29]; in proving it, rather than rely on control references not widely known in networking, we attempt a presentation as self-contained as possible, through the following proposition, also proved in the Appendix.

Proposition 2: Consider the unity feedback loop of Fig. 2, with $L(s) = F(s)\dfrac{I}{s}$. Suppose:

(i) $F(s)$ is analytic in $Re(s) > 0$ and bounded in $Re(s) \geq 0$.
(ii) $F(0)$ has strictly positive eigenvalues.
(iii) For all $\mu \in (0, 1]$ and $\omega \neq 0$, the point -1 is not an eigenvalue of $\mu L(j\omega)$.

Then the closed loop is stable.

In control jargon, the first two conditions imply that tuning down the loop gain by a small μ, there is negative feedback of enough rank to stabilize all the integrators; condition (iii) says that we can then increase μ up to unity without bifurcating into instability.

To apply this result to the loop transfer function in (13), we take $F(s) = \bar{R}_f(s)\mathcal{K}\bar{R}_b^T(s)\mathcal{C}$, which is easily seen to satisfy (i), and (ii) follows from the rank assumption on $\bar{R}$. To establish (iii), the key observation is the relationship

$$\bar{R}_b(s) = \bar{R}_f(-s)\mathrm{diag}(e^{-\tau_i s})$$

that follows from (3). This leads to the representation

$$L(j\omega) = \bar{R}_f(j\omega)\mathrm{diag}(\frac{\alpha_i x_{0i}}{M_i})\mathrm{diag}(\frac{e^{-\tau_i j\omega}}{\tau_i j\omega})\bar{R}_f(j\omega)^*\mathcal{C}, \tag{16}$$

where * denotes conjugate transpose. Isolating the factor

$$\Lambda(j\omega) := \mathrm{diag}(\lambda_i(j\omega)) = \mathrm{diag}\left(\frac{e^{-\tau_i j\omega}}{\tau_i j\omega}\right),$$

we see that it has eigenvalues on the curve Γ. The remainder of the proof involves showing that all the remaining factors produce nothing more than a convex combination and a scaling in these eigenvalues, and therefore can be prevented from reaching the critical point -1. This is done in the Appendix.

B. Nonlinear dual control and equilibrium structure

We have presented a linearized control law with some desirable stability properties. We now discuss how to embed such linear control laws in a global, nonlinear control scheme whose equilibrium would linearize as required.

The link control is simply (5) with our particular choice of γ_l, namely

$$\dot{p}_l = \begin{cases} \frac{y_l - c_{0l}}{c_{0l}}, & \text{if } p_l > 0 \text{ or } y_l > c_{0l}; \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

This gives our price units of time; indeed, for $c_{0l} = c_l$ this would correspond, by (7), to the *queueing delay* at the link (queue divided by capacity). Since we are working with a *virtual* capacity $c_{0l} < c_l$, we can interpret our price as the *virtual queueing delay* that the link would experience if its capacity were slightly lower.

For the sources, so far we have only characterized their linearization (15). For static source control laws as in (10), however, specifying its linearization at every equilibrium point essentially determines its nonlinear structure. Indeed, the linearization requirement (15) imposes that

$$\frac{\partial f_i}{\partial q_i} = -\frac{\alpha_i f_i}{M_i \tau_i},$$

for some $0 < \alpha_i < \pi/2$. Let us assume initially that α_i is constant. Then the above differential equation can be solved analytically, and gives the static source control law

$$x_i = f_i(q_i, \tau_i, M_i) = x_{\max,i}\, e^{-\frac{\alpha_i q_i}{M_i \tau_i}}. \tag{18}$$

Here $x_{\max,i}$ is a maximum rate parameter, which can vary for each source, and can also depend on M_i, τ_i (but not on q_i). This *exponential backoff* of source rates as a function of aggregate price can provide the desired control law, together with the link control in (17).

Remark: The RTT used in (18) could be the real-time measurement, or instead, it could be replaced by the fixed portion d_i of the delay. Both coincide locally around an equilibrium with

empty queues, but the latter option may be preferable because it avoids a more complex time-varying dynamics during a transient involving queueing delays. Later, we discuss practical ways for the source to estimate d_i.

The corresponding utility function (for which $f_i = (U_i')^{-1}$) is

$$U_i(x) = \frac{M_i\tau_i}{\alpha_i} x\left[1 - \log\left(\frac{x}{x_{\max,i}}\right)\right], \quad x \le x_{\max,i}.$$

We can achieve more freedom in the control law by letting the parameter α_i be a function of the operating point: in general, we would allow any mapping $x_i = f_i(q_i)$ that satisfies the differential inequality

$$0 \ge \frac{\partial f_i}{\partial q_i} \ge -\frac{\pi}{2}\frac{f_i}{M_i\tau_i}. \tag{19}$$

The essential requirement is that the slope of the source rate function (the "elasticity" in source demand) decreases with delay τ_i, and with the bound M_i on the number of bottlenecks.

So we find that in order to obtain this very general scalable stability theorem, some restrictions apply to the sources' demand curves (or their utility functions). This is undesirable from the point of view of our objective 3 in Section II-B; we would prefer to leave the utility functions completely up to the sources; in particular, to have the ability to allocate equilibrium rates independently of the RTT. We remark that parallel work in [32] has derived "primal" solutions with scalable stability and arbitrary utility functions, but where the link utilization objective is relaxed. Indeed, it appears that one must choose between the equilibrium conditions on either the source or on the link side, if one desires a scalable stability theorem. Below we show how this difficulty is overcome if we slightly relax our scalability requirement.

Finally, we emphasize that while the above implementations will behave as required around equilibrium, we have not proved global convergence to equilibrium in the nonlinear case. While our experiments with fluid simulations seem to support this fact, a mathematical proof is much more difficult; some results in the single link case are found in [34], [35]; see also [7] on global stability of the "primal" laws.

IV. A "PRIMAL-DUAL" LAW WITH RESOURCE ALLOCATION CONTROL

The reason we are getting restrictions on the equilibrium structure is that for static laws, the elasticity of the demand curve (the control gain at zero frequency) coincides with the high frequency gain, and is thus constrained by stability. To avoid this difficulty and thus allow for more flexibility in the rate assignment at equilibrium, we must decouple these two gains. This can only be done by adding dynamics at the sources, while still keeping the link dynamics, which guarantee network utilization. Thus we will have a "primal-dual" solution.

The simplest source control that has different gains at low and high frequencies is the first-order "lead-lag" compensator, given in the Laplace domain by

$$\delta x_i = -\frac{\kappa_i(s+z)}{s+\frac{z\kappa_i}{\nu_i}}\delta q_i. \tag{20}$$

Here the high frequency (as $s \to \infty$) gain κ_i is the same as in (15), "socially acceptable" from a dynamic perspective. The DC gain (at $s = 0$) $\nu_i = -f_i'(q_{i0})$ is the elasticity of source demand based on its own "selfish" demand curve $x_{i0} = f_i(q_{i0})$, that need no longer be of the form (18).

The remaining degree of freedom in the design is the choice of the *zero* z, which determines where the transition between "low" and "high" frequencies occurs. For reasons that have to do with the stability theorem below, it will be essential to fix this zero across all sources.

A. Local stability result

With the new local source control, we will proceed to study the linearized stability of the closed loop, generalizing the method of Theorem 1. We first write down the overall loop transfer function

$$L(s) = \bar{R}_f(s)\mathcal{K}(s)\bar{R}_b^T(s)\mathcal{C}\frac{I}{s}, \tag{21}$$

which is analogous to (13) except that now

$$\mathcal{K}(s) = \mathrm{diag}\big(\kappa_i V_i(s)\big), \quad \text{with} \quad V_i(s) = \frac{s+z}{s+\frac{z\kappa_i}{\nu_i}},$$

κ_i as in (15). The stability argument is based again on Proposition 2, the key step being once more the study of the eigenvalues of $\mu L(j\omega)$. We write

$$L(j\omega) = \bar{R}_f(j\omega)\mathrm{diag}(\frac{\alpha_i x_{0i}}{M_i})\Lambda(j\omega)\bar{R}_f(j\omega)^*\mathcal{C} \tag{22}$$

as in (16), except that now we have

$$\Lambda(j\omega) = \mathrm{diag}(\lambda_i(j\omega)) = \mathrm{diag}\left(\frac{e^{-\tau_i j\omega}}{\tau_i j\omega}V_i(j\omega)\right), \tag{23}$$

in other words we have added the lead-lag term $V_i(s)$ to the diagonal elements of $\Lambda(s)$. Since the remaining matrices are unchanged it will still be true (see the Appendix) that the eigenvalues of $L(j\omega)$ are convex combinations and scaling of these $\lambda_i(j\omega)$. So it remains to give conditions so that the convex combinations of the $\lambda_i(j\omega)$, which now include an extra lead-lag term, do not reach the critical point -1. Fig. 4 contains various Nyquist plots of $\lambda_i(j\omega)$, for τ_i ranging between 1ms and 1sec, and ratios ν_i/κ_i ranging between 0.1 and 1000. The value of z is fixed at 0.2.

A first comment is that here the plots do not coincide, as they did in the "scale-invariant" case of Section III; here only the high frequency portions coincide. Secondly, we note that there is not an obvious separation between the convex hull of these points and the critical point -1. One could think of obtaining convex separation through a slanted line; this however, would imply a lower limit $-\pi + \theta$, $\theta > 0$ on the phase of $\lambda_i(j\omega)$ at low frequencies, which in turn implies a limit on the lag-lead gain ratio ν_i/κ_i. This may be acceptable, but would not allow us to accommodate *arbitrary* utilities.

The alternative is to treat the low-frequency portion of the curves separately, ensuring for instance that they don't reach phase $-\pi$. This, however, implies a common notion of what "low-frequency" means, so that we are not operating in different portions of the curve for sources with different RTTs.

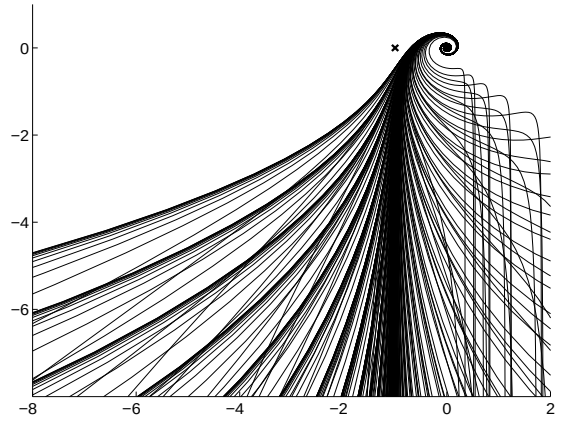

Fig. 4. Nyquist plots of $\lambda_i(j\omega)$, $z = 0.2$, various τ_i and ν_i/κ_i.

This can be obtained through a fixed bound $\bar{\tau}$ on the RTT, as follows.

Theorem 3: Assume that for every source i, $\tau_i \leq \bar{\tau}$. In the assumptions of Theorem 1 replace the source control by (20), with $\alpha_i \leq \alpha < \frac{\pi}{2}$ and $z = \frac{\eta}{\bar{\tau}}$. Then for a small enough $\eta \in (0, 1)$ depending only on α, the closed loop is linearly stable.

The proof is given in the Appendix.

B. Global nonlinear control

We now discuss how to embed our new linearized source control law in global nonlinear laws. The requirements are:

- The equilibrium matches the desired utility function, $U_i'(x_{0i}) = q_{0i}$, or equivalently the demand curve (4) for $f_i = (U_i)^{-1}$.
- The linearization is (20), with the zero z being fixed, independently of the operating point and the RTT.

We now present a nonlinear implementation that satisfies these conditions, which combines the structure of (18) with elements of the "primal" approach [20], [18], [32].

$$\tau_i \dot{\xi}_i = \beta_i (U_i'(x_i) - q_i), \tag{24}$$

$$x_i = x_{m,i} e^{(\xi_i - \frac{\alpha_i q_i}{M_i \tau_i})}. \tag{25}$$

Note that (25) corresponds exactly to the rate control law in (18), with the change that the parameter $x_{\max}$ is now varied exponentially as

$$x_{\max,i} = x_{m,i} e^{\xi_i},$$

with ξ_i as in (24). If β_i is small, the intuition is that the sources use (18) at fast time-scales, but slowly adapt their $x_{\max_i}$ to achieve an equilibrium rate that matches their utility function, as follows clearly from equation (24).

Remark: Here again, as in (18), it is often convenient to interpret τ_i as referring exclusively to the fixed portion d_i of the RTT.

We now find the linearization around equilibrium; the source subscript i is omitted for brevity. For increments $\xi = \xi_0 + \delta\xi$, $x = x_0 + \delta x$, $q = q_0 + \delta q$, we obtain the linearized equations:

$$\tau \delta\dot{\xi} = \beta\big(U''(x_0)\delta x - \delta q\big) = \beta\big(-\frac{\delta x}{\nu} - \delta q\big), \tag{26}$$

$$\delta x = x_0(\delta\xi - \frac{\alpha}{M\tau}\delta q) = x_0\delta\xi - \kappa\delta q. \tag{27}$$

Here we have used the fact that $U''(x_0) = \frac{1}{f'(q_0)} = -\frac{1}{\nu}$, and the expression (15) for κ . Some algebra in the Laplace domain leads to the transfer function

$$\delta x = -\kappa \left(\frac{s + \frac{\beta x_0}{\kappa\tau}}{s + \frac{\beta x_0}{\nu\tau}} \right) \delta q,$$

that is exactly of the form in (20) if we take

$$z = \frac{\beta x_0}{\kappa\tau} = \frac{\beta M}{\alpha}.$$

By choosing β, the zero of our lead-lag can be made independent of the operating point, or the delay, as desired.

We recapitulate the main result as follows.

Theorem 4: Consider the source control (24-25) where $U_i(x_i)$ is the source utility function, and the link control (17). At equilibrium, this system will satisfy the desired demand curve $x_{i0} = f_i(q_{i0})$, and the bottleneck links will satisfy $y_{0l} = c_{0l}$, with empty queues. Furthermore, under the rank assumption in Theorem 1, $\alpha_i < \frac{\pi}{2}$, and $z = \frac{\beta_i M_i}{\alpha_i}$ chosen as in Theorem 3, the equilibrium point will be locally stable.

We have thus satisfied all the equilibrium objectives set forth in Section II-B, and local stability. This was done for arbitrary networks, with the only restriction that an overall bound on the RTT had to be imposed.

Remark: Source laws (24-25) are not the only ones that satisfy our equilibrium and linearization objectives; we are aware of at least one alternative. Our preference for this version is based mainly on empirical observations of its global properties, and on its close relationship with the static law (18), for which there are some partial global stability results [34].

We conclude the section with a few remarks on the dynamic *performance* of the system, in particular its speed of convergence to equilibrium. Locally, the speed of response is dictated by the closed-loop poles, and it will be faster as one approaches the boundary of stability. How close to this boundary do we operate when using the parameter settings of Theorem 4? From the Nyquist argument one can easily see that the conditions are non-conservative in the case of a single bottleneck network shared by homogeneous sources, which is not an unrealistic scenario. Other aspects of the analysis can be more conservative: in particular regarding the scaling M_i, while it is needed to obtain a theorem for the worst-case network, in most examples stability occurs even if the number of bottlenecks is under-estimated.

A more important validation of performance is the *global* one, starting from a possibly far away initial condition, as would happen for instance after a change in routing or if a new source starts. This issue will inevitably depend on the utility function being used in (24), and (as with global stability) will be difficult to address other than by simulation.

Still, we can do the following approximate analysis to gain insight on the behavior of the control law as a new source

starts up from a very small rate. In particular the speed at which this rate grows will have a large impact on the time it takes to reach equilibrium. To analyze this, calculate from (24-25) the derivative of the rate,

$$\dot{x} = \frac{x}{\tau}\left(\beta U'(x) - \beta q - \frac{\alpha}{M}\dot{q}\right). \tag{28}$$

If the source were starting on an uncongested path the terms in q, $\dot{q}$ would disappear. This is also a good approximation for sources starting with a small rate on a steady-state, congested path. Indeed, in this case the marginal utility $U'(x)$ would be much larger than the price q, and also $\dot{q}$ would be small since the existing larger flows are in equilibrium. Therefore we can write the approximation

$$\dot{x} \approx \frac{\beta}{\tau} x U'(x) \tag{29}$$

for small x. We can use this to assess the performance of certain utility functions; for instance, the choice

$$U(x) = K \log(x) \tag{30}$$

which induces so-called *proportional fairness* in the equilibrium [20], makes

$$\dot{x} \approx \frac{\beta K}{\tau} \tag{31}$$

and therefore *linear* growth of the rates starting from zero. Instead, a utility function such that $U'(x)$ has a finite limit for $x \to 0$, will give initially an exponential increase of the rate.

V. A PACKET-LEVEL IMPLEMENTATION USING ECN MARKING

So far we have worked with the abstraction of the congestion control problem laid out in Section II-A. In this section we indicate how these ideas can transition to an actual packet-level protocol, as could be implemented in a real world network. For more details on the implementation aspect we refer to [30].

A first comment is that while we have assumed that source can control *rates*, in practice they adapt their congestion window w_i; its effect over the rate can be approximately described, over time-scales longer than the RTT, by the relationship

$$x_i \approx \frac{w_i}{\tau_i}. \tag{32}$$

Sources should set w_i so that the rate targets the desired "equation-based" value from (24-25), with a suitable time discretization interval T_s. To make the discussion more concrete, in this section we use the utility function $U_i(x_i) = K_i \log(x_i)$ from (30). A straightforward discretization of (24-25) could be

$$\xi_i(k) = \xi_i(k-1) + \beta_i \left(\frac{K_i}{w_i(k-1)} - \frac{q_i(k)}{d_i}\right) T_s, \tag{33}$$

$$w_i(k) = w_{m,i} e^{\ \xi_i(k) - \frac{\alpha_i q_i(k)}{M_i d_i}}. \tag{34}$$

An alternative discretization, that exploits (28) to avoid the complexity of computing the exponential, will be discussed in the following section.

To execute the above algorithm, sources must have access to their aggregate price (see below), their (minimum) RTT which can be measured through packet time-stamps, and the parameter M_i which must be assumed a priori. The latter is clearly the weakest point, although it could be argued that in the context of a network with uncongested backbones, most sources typically see few bottlenecks, so perhaps a value $M_i = 2$ would suffice.

Similarly, links can approximate (17) by a time discretization with interval $\tilde{T}_s$,

$$p(k) = \left[p(k-1) + \frac{y_l(k) - c_{0l}}{c_{0l}} \tilde{T}_s\right]^+. \tag{35}$$

Here $[\cdot]^+$ denotes $\max\{\cdot, 0\}$. Note that $y_l(k)\tilde{T}_s$ can be taken to be number of arrivals at the queue during the interval. Therefore the above operation can be performed with relatively small computational burden on the routers.

A. Marking and Estimation

The key remaining issue for the implementation of the above protocols is the communication of price signals from links back to sources, in an additive way across the source's route. In this section we explore the use of an Explicit Congestion Notification (ECN) bit to implement this feature. A natural way to obtain the additivity property is by Random Exponential Marking (REM, [1]), in which an ECN bit would be marked at each link l with probability

$$1 - \phi^{-p_l}$$

where $\phi > 1$ is a global constant. Assuming independence between links, the overall probability that a packet from source i gets marked is (see [1])

$$\mathcal{P}_i = 1 - \phi^{-q_i}, \tag{36}$$

and therefore q_i can be estimated from marking statistics. For example, a shift-register of the last N received marks can be maintained, the fraction of positive marks providing an estimate $\hat{\mathcal{P}}_i$ of the marking probability, from which an estimate $\hat{q}_i$ can be derived, and used in place of q_i in the source equations (33-34).

While simple in principle, two related issues are important to make this scheme practical:

1) The choice of a universal ϕ across the network means choosing a range of prices for which our estimation will be most accurate (i.e., where the marking probability is not too close to 0 or 1). For instance, choosing $\phi = 100$ implies the range of prices (in seconds) $[0.011, 0.65]$ corresponds to marking probabilities between 5% and 95%. In essence, ϕ selects a scale of prices, and source demand functions (4) should be tailored to operate with this "currency". In the simulations below, this will be taken into account in the choice of the constant K_i of our utility function.
2) An estimation based on a moving average of size N introduces an additional *delay* in the feedback loop, of approximately

$$\tau_{est} \approx \frac{N}{2w}\tau, \tag{37}$$

which is the time it takes to receive $\frac{N}{2}$ packets. This delay could compromise stability, a factor that can partly be addressed by choosing α away from the stability limit. Still, it is clear from (37) that one should avoid high estimation windows, so there is compromise between stability and accurate price estimation. Noise in price estimation will feed through to the congestion window by (34); this will not affect average rates, but it may nevertheless be undesirable. In the simulations below, we mitigate this noise by imposing caps on the window change at every sample time.

B. Simulation results

We implemented the preceding algorithms in the standard simulator ns-2 to validate their performance. The source estimates the price on each ACK arrival using the last N marks, and the round trip propagation delay from a minimum RTT; these are used to define an *expected* congestion window every T_s seconds based on (33-34). The actual congestion window is set to the expected window every ACK, but with a cap on the magnitude of the change. For more details, see [30]. The links run (35) to update price every $\tilde{T}_s$ seconds, and exponentially marks the ECN bit with base ϕ.

We used the following parameters in the simulation:

- $\phi = 10^6$, $N = 31$ for marking and estimation.
- $T_s = 10ms$, $\beta_i = 1.18$, $K_i = 50$, $M_i = 1$, $\alpha_i = 0.37$ at the sources.
- $\tilde{T}_s = 5ms$, $c_{0l} = 0.95c_l$ at the links. To focus on the control performance, we used large buffers to avoid packet drops.

The scenario of Fig. 5 tests the dynamics of our protocol when sudden changes in traffic demands take place. One-way long-lived traffic goes through a single bottleneck link with capacity of 2Gbps (250pkts/ms with mean packet size 1000bytes). It is shared at most by 512 ftp flows. The number of flows is doubled every 40 seconds, from 32, to 64, 128, 256, and finally to 512 flows. These groups of flows have round trip propagation delays of 40ms, 80ms, 120ms, 160ms and 200ms respectively. This scenario is designed to stress a high-capacity link with heterogeneous flows.

In reference to the results of Fig. 5, we note:

1) The source rates and the link prices (marking probability) track the expected equilibria when new sources activate. While there is noise in the price estimation, its impact is only significant in the very uncongested case, when prices are very low.
2) Fairness is achieved: at each equilibrium stage, the bandwidth is shared equally among sources despite their heterogeneous delays.
3) The queue is small (around 100 packets, less than 0.5 ms of queueing delay) almost all the time, both in transient and in equilibrium. The only (mild) queue overshoot is caused by the activation of 256 new flows in a short time.
4) After the startup transient of the first sources, link utilization remains always around the 95% target even when the traffic demand changes suddenly.

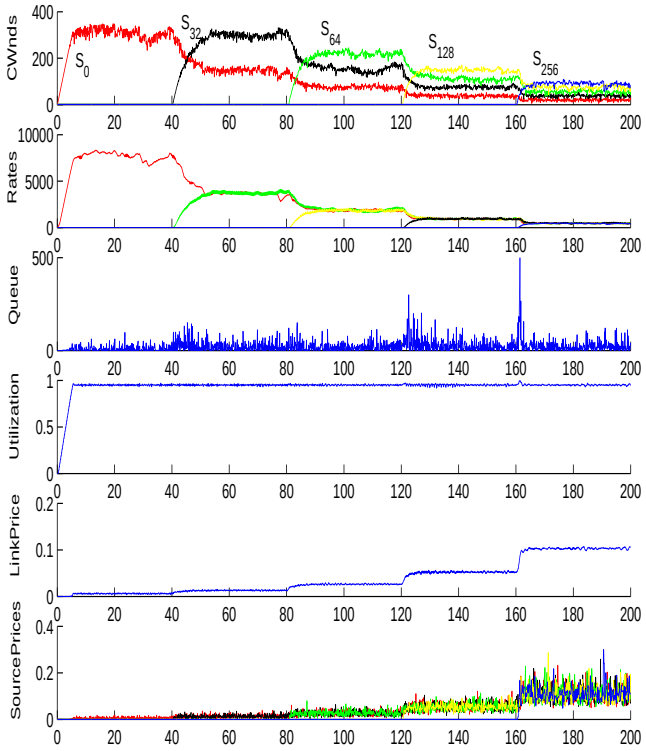

Fig. 5. Dynamic Performance of the ECN-based Protocol

Note that we are not using any "slow-start" phase in this protocol, we are running exclusively the algorithm described before. In fact, at the beginning of the simulation, when the price is small, the sources' rate grows approximately linearly, which can be explained by looking at equation (31). The slope of increase is approximately $\beta_i K_i/\tau_i$, so the utility function's parameter can be used to tune how aggressively the sources start-up, trading off speed with the risk of queue overshoots. If we wished an exponential increase in this initial stage, it may be advantageous to retain a slow-start phase, or use a different utility function, a factor we will explore in future work.

We have also performed extensive simulations of two-way traffic (when both data packets and ACKs share a congested link), and for situations where, instead of long-lived flows, we employ a "mice-elephant" mix of traffic. In particular, we included flow lengths drawn from a heavy-tailed distribution, which matches observed statistics in the Internet [33], [6]. Some of these simulation results are reported in [30]. We find that the protocol still keeps high utilization and small queues, and the elephants share the bandwidth fairly.

VI. A SOURCE-ONLY IMPLEMENTATION BASED ON QUEUEING DELAY

The protocol described above achieves close to full utilization of link capacity, while at the same time operating under essentially empty queues. These two properties can only be simultaneously achieved by some form of explicit congestion notification. In the absence of it, congestion can only be detected through some of its undesirable effects, such as the appearance of queueing delays. However, if one could

guarantee that these delays are moderate, it would perhaps be preferable to avoid the burden of ECN signaling on the routers.

In this regard, the fact that the prices in our protocol are virtual queueing delays, suggests the possibility of using *real* queuing delays as a price signal; this amounts to choosing $c_{0l} = c_l$ in the link equation (17). The advantage is that the sum of such delays over congested links can be estimated by sources by subtracting the minimum observed RTT (taken to be propagation delay) from the current RTT, avoiding any explicit signaling between links and sources. This is precisely the type of congestion feedback used in TCP-Vegas [3]. The question before us is to find a source protocol that can achieve (i) small equilibrium prices (delays), (ii) freedom of choice in resource allocation, and (iii) stability, working with queueing delay as a price signal.

Clearly, if we can assign arbitrary utility functions, we can use a constant factor in them to set the scale of equilibrium prices, similar to what we discussed in the context of marking. If, instead, we are "stuck" with a certainly family of utility functions, as in Section III, it may not be possible to control the equilibrium delay. For this reason we concentrate on extending the ideas of Section IV to the situation where queueing delays are allowed to appear in the network.

We first note that we must modify the source laws (24-25) if we wish to preserve dynamic structure under the current circumstances. In fact, before we had assumed that around equilibrium, the RTT τ_i was the same as the fixed (propagation/processing) delay d_i, and thus appeared only as a parameter in our linearization. That analysis is no longer valid here, because τ_i will be the *variable* quantity

$$\tau_i = d_i + q_i, \tag{38}$$

where q_i is the queueing delay observed by the source, and is also the price, nonzero in equilibrium. This leads us to propose the following alternative source laws:

$$\dot{\xi}_i = \frac{\beta_i}{(d_i + q_i)}(U_i'(x_i) - q_i), \tag{39}$$

$$x_i = x_{m,i} e^{\xi_i} \left(\frac{d_i}{d_i + q_i} \right)^{\frac{\alpha_i}{M_i}}. \tag{40}$$

Here, (39) is unchanged from (24), we have only made explicit the relationship (38) for the RTT. The change in (40) as compared to (25), is required to obtain the same input-output relationship between q_i and x_i, under the current circumstances. Indeed, taking derivatives in (40) and substituting with (39) we obtain (subindex dropped)

$$\begin{aligned} \dot{x} &= x_m e^{\xi} \dot{\xi} \left(\frac{d}{d+q} \right)^{\frac{\alpha}{M}} - x_m e^{\xi} \frac{\alpha}{M} \frac{d^{\frac{\alpha}{M}}}{(d+q)^{\frac{\alpha}{M}+1}} \dot{q} \\ &= x\dot{\xi} - x \frac{\alpha}{M(d+q)} \dot{q} \\ &= \frac{x}{d+q} \left(\beta U'(x) - \beta q - \frac{\alpha}{M} \dot{q} \right). \end{aligned} \tag{41}$$

The last equation is exactly the same as (28), again noting that $d + q = \tau$; in particular its linearization around equilibrium will still be (20), as desired.

Does this mean that our local stability theorem would hold for this protocol? Unfortunately there is another difficulty that arises from queueing delays; namely, that the network equations (1) and (2) become time-varying. In fact, an expression such as

$$x_i(t - \tau_{li}^f)$$

is difficult to handle, and even to interpret, if τ_{li}^f depends on time, and further if it does so through other state variables $p_l(t)$. In particular, given the time-varying nature of this system we cannot rigorously employ Laplace transforms, which were the basis of our linear theory. At most, this analysis can be considered as an approximation to the linearized dynamics, where we consider only variations $\tau_i(t)$ in the *dependent* variables (e.g. the scaling of source rates), but not in the *independent* (time) variable. This kind of approximation has been used successfully [9], [26] to analyze TCP-Reno dynamics but has not been rigorously justified.

If we adopt this approximation, we could write the expressions (8-9), where now the matrices $\bar{R}_f(s)$ and $\bar{R}_b(s)$ are defined in terms of the *equilibrium* forward and backward delays, including queueing. The resulting overall system obtained from the source laws (39-40) and simple queues (7) at the links, is indeed locally stable under similar assumptions on the parameters. Thus there is hope that this protocol would behave satisfactorily, but we must rely (more so than before) on empirical simulations to validate this fact.

A. Packet implementation and simulation results

In this case, links would be passive and produce queueing delays in the natural way. The only consideration here is that we assume the buffering is enough to allow operation without saturation. This, again, relates to the choice of utility function parameters.

As for the source window dynamics, (39-40) could be discretized directly, analogously to (33-34), however we present here an alternative discretization based on (41), which has lower complexity[3]. For the utility function under consideration, rewrite (41) as

$$\dot{x} = \frac{\beta K}{\tau} - x \left(\frac{\beta q}{\tau} + \frac{\alpha}{M\tau} \dot{q} \right), \tag{42}$$

that is approximated by the following window update performed every T_s seconds:

$$\begin{aligned} w(k+1) &= \beta K T_s \\ &\quad + w(k) \left[1 - \left(\beta T_s + \frac{\alpha}{M} \right) \frac{q(k)}{\tau(k)} + \frac{\alpha}{M} \frac{q(k-1)}{\tau(k)} \right]. \end{aligned} \tag{43}$$

The queueing delay value in the above equation would be estimated as in TCP Vegas [3] (RTT $-$ minimum RTT), but here the window dynamics is chosen to provide the stability guarantees. For other work on stabilizing Vegas, see [4].

Fig. 6 uses the same scenario and required parameter values as in Section V-B. The simulation shows fast response to the traffic demand and stability. Furthermore, the windows are extremely smooth as well as the queues due to the accurate estimation of the price, i.e., the queueing delay. This, and the

[3]This version could also be applied to the ECN case via (28), however we have found that in a noisy environment it can lead to bias, inducing unfairness.

lack of complexity at routers, are interesting advantages of this protocol with respect to the ECN version. There are, however, drawbacks: one, that a certain amount of queueing delay must be tolerated here. While parameters (e.g. K_i in the utility function) can be tuned to make it small, there is a tradeoff with the speed of response of the system. Another issue that appears is some unfairness, caused by sources joining the network later that overestimate the propagation delay, and thus underestimate price, taking up a larger share of the bandwidth.

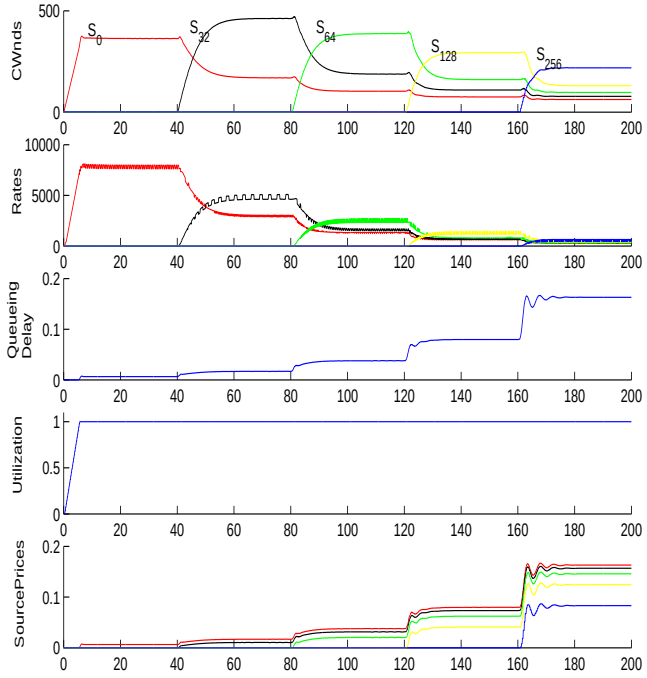

Fig. 6. Dynamics of protocol based on queueing delay

VII. Conclusion

A congestion avoidance method that can achieve high utilization, small queueing delay, freedom from oscillations and fairness in bandwidth allocation has been a major objective of networking research in recent years. A fundamental question in this regard is how far we can go in achieving these objectives within the existing end-to-end philosophy of the Internet.

Our results show that if the fairness aspect is expressed in terms of utility functions [21], local regulation around these desirable equilibrium points can be achieved through a very minimal feedback mechanism: a scalar price signal that reflects the aggregate congestion level of the traversed links for each source. Furthermore, convergence results can be made independent of network topology, routing, parameters, and delay except for a commonly agreed bound. We are currently working [34], [35] on a better understanding of the nonlinear dynamics, which has significant impact on the speed of the control away from equilibrium.

We have further demonstrated a practical version of the protocol based on ECN marking, that appears to successfully approximate these objectives in high capacity networks where current protocols exhibit limitations. Compared to other proposed solutions (e.g. [19]), this ECN version represents a minor change in the protocol implementations at routers and end-systems. Still, it would be clearly preferable to have to upgrade only one end of things; this motivated us to consider the implementation based on queueing delay, similar to TCP Vegas, which appears capable of delivering most of the benefits with no active participation of network routers. Based on our preliminary success in simulations, we are currently pursuing experimental deployment of these kinds of protocols [8]. Part of this effort involves testing the coexistence of such protocols with deployed versions of TCP.

Appendix

Proof of Proposition 2

A first condition for internal stability in Fig. 2 is that there must be no unstable pole-zero cancellations when computing $L(s)$; this is true because $F(s)$ is stable and $F(0)$ is invertible from assumption (ii), so there is no cancellation at $s = 0$. Stability now reduces to showing that

$$(I + L(s))^{-1} = s\big(sI + F(s)\big)^{-1}$$

is analytic and bounded in $Re(s) \geq 0$, or equivalently that $\det(sI + F(s))$ has no roots in this region. We will show, in fact, that $\varphi_\mu(s) := \det(sI + \mu F(s))$ has no roots in $Re(s) \geq 0$ for any $\mu \in (0, 1]$.

First, using (ii) and continuity, select $\epsilon > 0$ such that for $|s| \leq \epsilon$, and $Re(s) \geq 0$, we have $Re(\mathrm{eig}(F(s))) > 0$. Clearly, for such s there can be no roots of $\varphi_\mu(s)$, otherwise this would give an eigenvalue $-s/\mu$ of $F(s)$, with non-positive real part.

It remains to consider the region $|s| > \epsilon$, $Re(s) \geq 0$. If C is a bound of $\|F(s)\|$ (obtained from hypothesis (i)), then there can be no roots of $\varphi_\mu(s)$ for $0 < \mu < \epsilon/K$, because here

$$\left\| \frac{\mu F(s)}{s} \right\| < \frac{\mu K}{\epsilon} < 1,$$

and therefore $I + \mu F(s)/s$ is invertible. We now increase μ from this range up to 1. If a root of $\varphi_\mu(s)$ ever goes into $Re(s) \geq 0$, by continuity there must exist $\mu \leq 1$ for which the root is in $Re(s) = 0$. Since $s = 0$ is never a root, we have

$$\det(j\omega I + \mu F(j\omega)) = 0 \quad \text{for some} \quad \omega \neq 0.$$

But then $-1 \in \mathrm{eig}(\mu L(j\omega))$, contradicting (iii).

■

Proof of Theorem 1

What remains is to establish that the loop transfer function in (13) satisfies condition (iii) in Proposition 2. Referring back to (16), we write the new expression

$$L(j\omega) = \bar{R}_f(j\omega) X_0 \mathcal{A}\mathcal{M}\Lambda(j\omega) \bar{R}_f(j\omega)^* \mathcal{C}, \qquad (44)$$

where we recall that

$$\Lambda(j\omega) := \mathrm{diag}(\lambda_i(j\omega)) = \mathrm{diag}\left(\frac{e^{-\tau_i j\omega}}{\tau_i j\omega}\right),$$

and we have introduced the new notation

$$X_0 = \mathrm{diag}(x_{0i}), \quad \mathcal{A} = \mathrm{diag}(\alpha_i), \quad \mathcal{M} = \mathrm{diag}(\frac{1}{M_i}).$$

We now use the fact that nonzero eigenvalues are invariant under commutation, and that many of the factors in (44) are diagonal, to conclude that

$$-1 \in \text{eig}\Big(\mu L(j\omega)\Big) \Longleftrightarrow -1 \in \text{eig}\Big(P(j\omega)\Lambda(j\omega)\Big), \quad (45)$$

where the matrix $P(j\omega) \geq 0$ is defined as

$$P(j\omega) := \mu \mathcal{M}^{\frac{1}{2}} \mathcal{A}^{\frac{1}{2}} X_0^{\frac{1}{2}} \bar{R}_f(j\omega)^* \ \mathcal{C} \ \bar{R}_f(j\omega) X_0^{\frac{1}{2}} \mathcal{A}^{\frac{1}{2}} \mathcal{M}^{\frac{1}{2}}. \quad (46)$$

Claim: The spectral radius $\rho(P) < \frac{\pi}{2}$. To establish this, write (note $\mu \leq 1$)

$$\begin{aligned} \rho(P) &= \rho\Big(\mu \mathcal{M} \bar{R}_f(j\omega)^* \ \mathcal{C} \ \bar{R}_f(j\omega) X_0 \mathcal{A}\Big) \\ &\leq \|\mathcal{M} \bar{R}_f(j\omega)^*\| \cdot \|\mathcal{C} \ \bar{R}_f(j\omega) X_0\| \cdot \|\mathcal{A}\|. \end{aligned}$$

Any induced norm will do, but if we use the l_∞-induced (max-row-sum) norm, we find that

$$\begin{aligned} \|\mathcal{C} \bar{R}_f(j\omega) X_0\|_{\infty - ind} &= \max_l \frac{1}{c_{0l}} \sum_{i \ \text{uses} \ l} |e^{-\tau_{i,l}^f j\omega} x_{0i}| \\ &= \max_l \frac{1}{c_{0l}} \sum_{i \ \text{uses} \ l} x_{0i} = 1; \end{aligned}$$

note we are dealing with bottlenecks. Also $\|\mathcal{M} \bar{R}_f^*\| \leq 1$, because each row of this matrix contains at most M_i nonzero elements of magnitude $1/M_i$. Finally, $\|\mathcal{A}\| < \frac{\pi}{2}$ by hypothesis.

So $\rho(P) < \frac{\pi}{2}$ as claimed. This claim can be used as in [29] to show directly by contradiction that -1 is not an eigenvalue of $\mu L(j\omega)$. It is more concise however to invoke the following Lemma from [31] that elegantly characterizes the eigenvalues of the product of a positive and a diagonal matrix as in (45).

Lemma 5 (Vinnicombe): Let $P = P^* \geq 0$ and $\Lambda = \text{diag}(\lambda_i)$ be $n \times n$ matrices, then the eigenvalues of $P\Lambda$ belong to the convex hull of $\{0, \lambda_1, \ldots, \lambda_n\}$, scaled by the spectral radius $\rho(P)$.

Here, the points $\{0, \lambda_1, \ldots, \lambda_n\}$ all belong to the curve Γ on Fig. 3; its convex hull intersects the negative real axis in the segment $[-\frac{2}{\pi}, 0]$. Our above claim implies that scaling by $\rho(P)$ one cannot reach the critical point -1. Thus we establish condition (iii) in Proposition 2, and through it we conclude the proof of Theorem 1.

■

Proof of Theorem 3

As discussed in Section IV-A, we parallel the argument for Theorem 1, based on Proposition 2. Once again, the problem reduces to establishing that $-1 \notin \text{eig}(P(j\omega)\Lambda(j\omega))$, where $P(j\omega)$ is unchanged from (46), but now the diagonal elements of Λ are of the form

$$\lambda_i(j\omega) = \frac{e^{-\tau_i j\omega}}{\tau_i j\omega} V_i(j\omega), \qquad V_i(s) = \frac{s+z}{s + \frac{z\kappa_i}{\nu_i}}.$$

Invoking Lemma 5, we must study the convex combinations of these new λ_i's; this we do by breaking the analysis in two frequency regions, and using the hypothesis $z = \frac{\eta}{\bar{\tau}}$.

- For frequencies $\omega \geq \frac{1}{\bar{\tau}}$, we quantify the extra gain and phase introduced by $V_i(j\omega)$:

$$\begin{aligned} |V_i(j\omega)| &\leq \left|\frac{j\omega + z}{j\omega}\right| = \sqrt{1 + \frac{z^2}{\omega^2}} \leq \sqrt{1+\eta^2}, \\ \text{phase}\,(V_i(j\omega)) &\geq \text{phase}\left(\frac{j\omega+z}{j\omega}\right) \\ &= -\arctan(\frac{z}{\omega}) \geq -\arctan(\eta). \end{aligned}$$

  Since the first factor in $\lambda_i(j\omega)$ belongs to our familiar curve Γ (solid line in Fig. 7), we find that $\lambda_i(j\omega)$ will always lie below the perturbed curve

$$\Gamma_\eta := \sqrt{1+\eta^2} e^{-j \arctan(\eta)} \Gamma$$

  (a slight clockwise rotation and expansion of Γ), depicted by dashed lines in Fig. 7. Let $-g(\eta)$ denoted the first

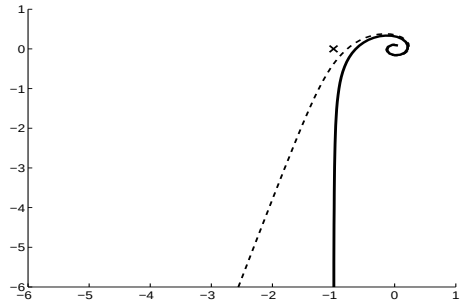

Fig. 7. Plots of Γ (solid) and Γ_η (dashed)

  point where Γ_η intersects the negative real axis. Since $g(0) = 2/\pi$, we can choose η small enough so that $g(\eta)\alpha < 1$ (recall that $\alpha < \pi/2$; how small η needs to be depends only on the "robustness margin" between α and $\pi/2$). With this assumption, we see that convex combinations of points below the curve Γ_η, scaled up to α, cannot reach the critical point -1. But, analogously to the previous theorem, we have that $\rho(P) \leq \alpha$; so Lemma 5 implies the critical point will not be reached in this frequency region.
- For frequencies $\omega \in (0, \frac{1}{\bar{\tau}})$, we will argue that $\lambda_i(j\omega)$ is always in the lower half-plane (negative imaginary part), and hence again one cannot obtain the critical point by convex combination and scaling. To see this, compute

$$\begin{aligned} \text{phase}(\lambda_i(j\omega)) &= -\frac{\pi}{2} - \tau_i\omega + \text{phase}(V_i(j\omega)) \\ &> -\pi - \tau_i\omega + \arctan(\frac{\omega}{z}) \\ &\geq -\pi - \bar{\tau}\omega + \arctan(\frac{\bar{\tau}\omega}{\eta}). \end{aligned}$$

  Thus it suffices to show that for $\omega \in (0, \frac{1}{\bar{\tau}})$,

$$\arctan(\frac{\bar{\tau}\omega}{\eta}) > \bar{\tau}\omega,$$

  or equivalently $\eta < \frac{\bar{\tau}\omega}{\tan(\bar{\tau}\omega)}$. The right hand-side is decreasing in $\bar{\tau}\omega < 1$, so it suffices to choose $\eta < \frac{1}{\tan(1)} \approx 0.64$. ■

ACKNOWLEDGEMENT

This work was influenced by many discussions that took place at the Spring '02 UCLA-IPAM workshop on Large-Scale Communication Networks; we are grateful to Frank Kelly, Tom Kelly, Srisankar Kunniyur, R. Srikant and Glenn Vinnicombe for this fertile interaction.

REFERENCES

[1] S. Athuraliya, V. H. Li, S. H. Low, and Q. Yin, "REM: active queue management", *IEEE Network*, vol. 15, no. 3, pp.48-53, 2001.

[2] D. Berstekas and R. Gallager, *Data Networks*, Prentice-Hall 1992.

[3] L. S. Brakmo and L. Peterson, "TCP Vegas: end to end congestion avoidance on a global Internet", *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 8, October 1995.

[4] D. H. Choe and S. H. Low, "Stabilized Vegas", *Proceedings 2003 IEEE Infocom.*.

[5] D.D. Clark, "The design philosophy of the DARPA Internet protocols", *Proc. ACM SIGCOMM '88, in: ACM Computer Communication Reviews*, Vol. 18, No 4., pp. 106-114, 1988.

[6] M. E. Crovella and A. Bestavros. "Self–similarity in World Wide Web traffic: evidence and possible causes." *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997.

[7] S. Deb, R. Srikant,"Global Stability of Congestion Controllers for the Internet," *IEEE Trans. on Automatic Control*, Vol. 48, No 6., pp. 1055-1059, December 2003.

[8] FAST Project, http://netlab.caltech.edu/FAST/.

[9] C. Hollot, V. Misra, D. Towsley, and W-B Gong, "A control theoretic analysis of RED", in *Proc. IEEE Infocom*, April 2001.

[10] V. Firoiu and M. Borden, "A study of active queue management for congestion control," in *Proc. IEEE Infocom*, March 2000.

[11] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance", *IEEE/ACM Trans. on Networking*, vol. 1, no. 4, pp. 397-413, August 1993.

[12] S. Floyd, *"HighSpeed TCP for large congestion windows"*, Internet draft, June 2002. http://www.ietf.org/internet-drafts/draft-floyd-tcp-highspeed-00.txt

[13] R.J. Gibbens and F.P.Kelly, "Resource pricing and the evolution of congestion control", *Automatica* 35 (1999), 1969-1985.

[14] R.J. Gibbens and F.P.Kelly, "Distributed connection acceptance control for a connectionless network", *Proc. 16th Teletraffic Congress*, Edinburgh, UK, June 1999.

[15] F. Baccelli and D. Hong, "Interaction of TCP flows as billiards", Proc. 2003 IEEE Infocom.

[16] V. Jacobson, "Congestion avoidance and control", *Proc. SIGCOMM'88, ACM*.

[17] D.M. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks", *Comp. Networks and ISDN Sys.*, 17, pp. 1-14, 1989.

[18] R. Johari and D. Tan, "End-to-end congestion control for the Internet: delays and stability", *IEEE/ACM Transactions on Networking* 9(2001) 818-832.

[19] D. Katabi, M. Handley, and C. Rohrs, "Internet Congestion Control for High Bandwidth-Delay Product Networks." *Proc. ACM Sigcomm*, Pittsburgh, PA, August 2002.

[20] F. P. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: Shadow prices, proportional fairness and stability", *Jour. Oper. Res. Society*, vol. 49, no.3, pp 237-252, March 1998.

[21] F.P. Kelly, "Models for a self-managed Internet", *Philosophical Transactions of the Royal Soceity*, 358(2000) 2335-2348.

[22] S. Kunniyur and R. Srikant, "A time-scale decomposition approach to adaptive ECN marking", *IEEE Trans. on Automatic Control*, Vol. 47, No6., pp. 882-894, June 2002.

[23] S. Kunniyur and R. Srikant, "Stable, scalable, fair congestion control and AQM schemes that achieve High Utilization in the Internet", *IEEE Trans. on Automatic Control*, Vol. 48, No11., pp. 2024-2029, December 2003.

[24] B. Kuo and F. Golnaraghi, *Automatic Control Systems*, 8th edition, Wiley, NY 2003.

[25] S. H. Low and D. E. Lapsley, "Optimization flow control, I: basic algorithm and convergence", *IEEE/ACM Transactions on Networking*, vol.7, no.6,pp861-874, December 1999.

[26] S. H. Low, F. Paganini, J. Wang, S. A. Adlakha, and J. C. Doyle, "Dynamics of TCP/RED and a scalable control", *Proceedings 2002 IEEE Infocom.*.

[27] S. H. Low, F. Paganini, J. C. Doyle, "Internet congestion control", *IEEE Control Systems Magazine*, February 2002.

[28] Massoulie, L., "Stability of Distributed Congestion Control with Heterogeneous Feedback Delays", *IEEE Trans. on Aut. Control*, vol, 47, pp. 895-902.

[29] F. Paganini, J. C. Doyle,and Steven H. Low, "Scalable laws for stable network congestion control", in *Proc. IEEE Conference on Decision & Control*, Orlando, FL, 2001.

[30] F. Paganini, Z. Wang, S. Low and J. Doyle, "A new TCP/AQM for Stable Operation in Fast Networks", *Proceedings 2003 IEEE Infocom.*.

[31] G.Vinnicombe, "On the stability of end-to-end congestion control for the Internet", Tech. Rep., Cambridge University, CUED/F-INFENG/TR.398, December 2000.

[32] G. Vinnicombe, "On the stability of networks operating TCP-like congestion control", Proceedings 2002 IFAC World Congress.

[33] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. "Self–similarity through high variability: statistical analysis of Ethernet LAN traffic at the source level." *IEEE/ACM Transactions on Networking*, 5(1):71–86, 1997.

[34] Z. Wang and F. Paganini, "Global Stability with Time Delay in Network Congestion Control", *Proc. IEEE Conf. on Decision and Control*,, Las Vegas, NV, 2002.

[35] Z. Wang and F. Paganini, "Boundedness and Global Stability of a Nonlinear Congestion Control with Delays", submitted to *IEEE Trans. on Automatic Control*.