Achieving 100% Throughput in TCP/AQM Under Aggressive Packet Marking With Small Buffer

Do Young Eun, *Member, IEEE*, and Xinbing Wang, *Member, IEEE*

***Abstract*—We consider a TCP/AQM system with large link capacity (NC) shared by many flows. The traditional rule-of-thumb suggests that the buffer size be chosen in proportion to the number of flows (N) for full link utilization, while recent research outcomes show that $O(\sqrt{N})$ buffer sizing is sufficient for high utilization and $O(1)$ buffer sizing makes the system stable at the cost of reduced link utilization. In this paper, we consider a system where the Active Queue Management (AQM) is scaled as $O(N^{\alpha})$ with a buffer of size $O(N^{\beta})$ $(0 < \alpha < \beta < 0.5)$. By capturing randomness both in packet arrivals and in packet markings, we develop a doubly-stochastic model for a TCP/AQM system with many flows. We prove that, under such a scale, the system always performs well in the sense that the link utilization goes to 100% and the loss ratio decreases to zero as the system size N increases. Our results assert that the system enjoys benefit of largeness with no tradeoff between full link utilization, zero packet loss, and small buffer size, at least asymptotically. This is in stark contrast to existing results showing that there always exists a tradeoff between full link utilization and the required buffer size. Extensive ns-2 simulation results under various configurations also confirm our theoretical findings. Our study illustrates that blind application of fluid modeling may result in strange results and exemplifies the importance of choosing a right modeling approach for different scaling regimes.**

***Index Terms*—Router buffer sizing, small buffer, stochastic modeling, transmission control protocol.**

I. Introduction

IN THE current Internet, TCP congestion control is responsible for carrying about 90% of the bytes of total traffic generated in the network. Such a congestion control algorithm consists of a network and a source algorithm that are tightly coupled with each other. The network algorithm, or the Active Queue Management (AQM) usually acting on a router, detects an onset of congestion and governs how to generate and update *congestion signal* based on the information collected at the router, e.g., input traffic rate, queue size, delay, etc., and then notify the sender, hoping that the sender will react appropriately to help alleviate the congestion and to attain best network performance. On the other hand, the source algorithm that operates at the edge of the network (end-user), dictates how each sender should change its transmission rate or window size in response to the congestion signal (packet drop, delay, or Explicit Congestion Notification (ECN) marks [1]) from the network.

As the number of flows and the size of link capacity in the network continue to grow, the analysis and design of such a large TCP/AQM system are becoming increasingly difficult and involved. Among other issues, the question of buffer sizing has recently received much attention in the literature [2]–[8]. The question is: given the number of TCP flows (N) and the size of link capacity (NC), how do we choose the buffer size $B(N)$? Under the linear buffer sizing $B(N) = O(N)$,[1] it is well known that a "stable" system ensures the convergence of a normalized queue-length to a nonzero value (nonzero queueing delay) and thus achieves 100% link utilization [9], [3], [10], [11], while the square-root buffer sizing $(B(N) = O(\sqrt{N})$ turns out to be sufficient to achieve reasonably high link utilization when N is large [2]. More recently, it has been suggested that a very small buffer of size $(B(N) = O(1))$ would be enough and in fact make the system stable at the cost of reduced link utilization [11], [5], [6], [8]. Clearly, different objectives give rise to different guidelines on buffer sizing, and there exists a tradeoff between full link utilization and the amount of required buffer space.

On the other hand, for the design of TCP/AQM schemes associated with the chosen buffer size, the "fluid modeling" of TCP/AQM congestion control and its stability analysis have proven extremely powerful and versatile. Still, different scaling regimes for AQM schemes and buffer sizes often lead to different types of fluid models. For example, when there are N flows with capacity NC and linear scaling $(O(N))$ is employed for the buffer size and AQM (e.g., all the buffer thresholds for packet marking are $O(N)$), it has been shown that the system dynamics can be described by the *normalized* version of the system [12], [13] via the law-of-large-numbers type of arguments. In this case, several stability criteria have been obtained in terms of parameters of the normalized system [14], [15], [10], where the packet marking is based on the normalized queue-length and the stability of the system refers to the convergence of this normalized queue-length in the steady-state (thus resulting in 100% link utilization). In contrast, when the buffer size and the scaling for AQM are both chosen as $O(1)$, it turns out that the system is better modeled by explicitly capturing the random packet arrivals within each RTT (rate-based AQMs) [16], [11], [5], [17], where the packet marking is based on the normalized rate into the queue and the stability of the system now refers to the convergence of this normalized rate (which is less than

Manuscript received November 22, 2005; revised August 28, 2006, and April 3, 2007; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor R. Srikant. This work was supported in part by the National Science Foundation through CAREER Award CNS-0545893.

D. Y. Eun is with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695-7911 USA (e-mail: dyeun@ncsu.edu).

X. Wang is with the Department of Electrical Engineering, Shanghai Jiaotong University, Shanghai 200240, China (e-mail: xwang8@sjtu.edu.cn).

Digital Object Identifier 10.1109/TNET.2007.904000

[1] Throughout this paper, we write $f(x) = O(g(x))$ when $0 < \liminf_{x\to\infty} \frac{f(x)}{(g(x)} \leq \limsup_{x\to\infty} \frac{f(x)}{g(x)} < \infty$.

the normalized capacity, thus leading to a reduced link utilization). An "intermediate" scaling of $O(N^{\gamma})$ with $0 < \gamma < 1$ for AQMs and the buffer size was also considered in [5]. However, the attention was paid mainly to the stability of the system, as opposed to the performance in terms of the link utilization and the queue-length distribution in the steady-state.

In this paper, we focus on TCP/AQM systems with ECN marking under a "sub-square-root" scaling and investigate the system performance in terms of the link utilization and the queue-length distribution in the steady-state. Specifically, we scale the AQM and the buffer size as $O(N^{\alpha})$ with $0 < \alpha < 0.5$ (more "aggressive" than $O(\sqrt{N})$) and analyze the system behavior as N becomes large. By explicitly taking into account the randomness both in packet arrival time and the number of packets itself, we develop a "doubly-stochastic" Markovian model in which the aggregate arrival to the queue is modeled by a conditional Poisson process—a Poisson process with stochastic rate where the stochastic rate is determined by the dynamics of random packet markings in the previous RTT. We then analyze this doubly-stochastic model under the proposed scale in the steady-state and show that, as N increases: 1) the link utilization converges to 100% and 2) the queue-length distribution is concentrated right on "target," leading to zero loss probability. Thus our results show that the system enjoys "benefit of largeness" under $O(N^{\alpha})$ scale in that there is *no tradeoff* between 100% link utilization, zero loss probability at the queue, and smaller buffer size, at least asymptotically. This is in sharp contrast to other existing results in the literature, which all predict some tradeoffs between the buffer size and the link utilization. We also provide extensive simulation results under various settings to confirm all our theoretical findings.

The rest of the paper is organized as follows. In Section II, we first provide background on the issue of buffer sizing at routers and propose our scales for AQM and the buffer sizing. We also illustrate the difficulty of finding a "right" fluid model for the system under the aggressive scale. In Section III, we develop a doubly-stochastic framework to capture the randomness both in packet marking and time of arrivals and describe our modeling details. In Section IV, we prove our main results showing that all the performance metrics are not impaired under the proposed scale asymptotically. In Section V, we provide extensive ns-2 simulations to verify our findings. In Section VI, we illustrate the importance of capturing the randomness in packet arrivals under the aggressive scale and discuss existing results for the proposed scale. Finally, we conclude in Section VII.

II. Preliminaries

A. Scaling Buffers and AQMs Inside a Network

In TCP congestion control, one of the long-held rules-of-thumb in network design is that the buffer size at bottleneck links should be proportional to the bandwidth-delay product [9], [3]. In a large network setting where there are N flows with link capacity NC, this rule-of-thumb suggests $O(N)$ for buffer sizing. Associated with this buffer sizing is the linear scaling for AQM, in which the marking function $p^N(x) \in [0,1]$, or the probability of packet marking when the queue-length is x, is scaled *linearly*, i.e., for all N and x, $p^N(Nx) = p(x)$ for some function p, where the superscript N in the marking function $p^N(x)$ means that there are N flows with capacity NC [18], [10], [13], [19], [15]. Note that the linear scaling means that packets are marked based on the normalized queue-length.

Under the drop-tail AQM policy with many flows, it is shown [2] that one can do much better than this traditional rule-of-thumb for buffer sizing. *Empirical observation* reveals that when N is large, each user's window size $W_i (i = 1, 2, \ldots, N)$ becomes independent. By appealing to the Central Limit Theorem, their sum will behave like a Gaussian random variable with mean equal to the size of the "pipe" $(\mathrm{NC} \times \mathrm{RTT})$ and with standard deviation $O(\sqrt{N})$. Hence, by choosing the buffer size on the order of $\sqrt{N}$ to absorb typical fluctuations, one can still achieve high utilization and thus save huge cost for buffers implemented in high-speed core routers. Further, quite recently, far smaller buffer sizing of $B(N) = O(1)$ has been proposed in [5], [6], [8], all of which however result in a reduced link utilization (bounded away from 1).

This observation leads us to pose the following question: can we do better than $O(\sqrt{N})$, i.e., can we put buffers of size $O(N^{\beta})$ with $\beta \in (0, 0.5)$ instead, while maintaining full link utilization and low packet loss? The only available result in the literature using this range of scale is in [5], where it was suggested that the system under $O(N^{\gamma})$ $(0 < \gamma < 1)$ scale would promote instability for very large N (thus undesirable). Still, it lacks any performance investigation in terms of some possible tradeoff between the link utilization and the queue-length behavior under such a scale.

B. Aggressive Packet Marking

In this paper, we explore any possibility of using buffer sizes of $O(N^{\beta})$ and AQM schemes with $O(N^{\alpha})$ scaling, where $0 < \alpha < \beta < 0.5$. By this scaling of TCP/AQM, we mean that the buffer size is on the order of N^{β} $(B(N) = O(N^{\beta}))$ and, for AQM schemes, there exists a function p such that $p^N(N^{\alpha}x) = p(x)$ for all N and $x \geq 0$ (see Fig. 1). Note that this can be interpreted as a more *aggressive* marking than the case of $\alpha = \beta = 1$ (linear scale) since the marking probability is much higher for the same queue-length.[2] We note that difference in marking scales simply means different parameter setup at the routers, which can be easily configured as desired. Moreover, the scaling of marking functions, or AQM schemes in general, is directly related to the issues of network design over a long time scale. For example, if the number of connections (or customers) were to increase by ten times, then a natural question to ask is: what percentages of additional capacities or buffer sizes are needed at routers to provide the same level of performance to each user? Under the linear scale $(\alpha = \beta = 1)$, we would have to increase by 10 times the buffer size as well as all the parameters associated with the AQM. If our aggressive scale is to be used, the buffer size and the AQM parameters should be increased very little (e.g., $10^{0.3} \approx 2$ and $10^{0.2} \approx 1.6$ when $\alpha = 0.2$ and $\beta = 0.3$).

Fig. 1 shows some examples of marking functions with the scale of N^{α}. Note that all the thresholds for the marking are on

[2] Our scale is even more aggressive than the case of $\alpha = 0.5$.

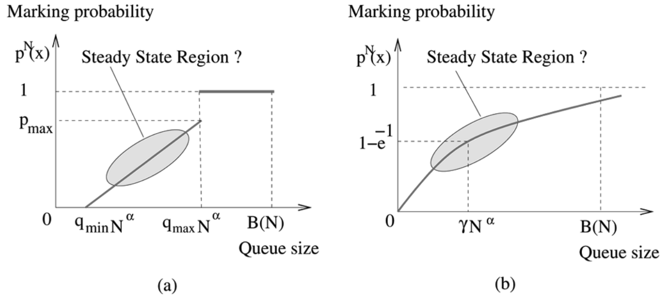


Fig. 1. Examples of marking functions with aggressive marking scale of N^{α}. (a) Linear marking: RED type. (b) Exponential marking: REM type.

the order of N^{α}. Suppose that the system performs as desired in the steady-state,[3] i.e., the queue-length will be $O(N^{\alpha})$ with high probability. Then, we expect that the system will achieve almost 100% utilization while packet loss probability with a buffer of size $O(N^{\beta})$ $(\beta > \alpha)$ will be very small large N. Our *ns*-2 simulation results show that this is indeed the case (see Section V for details.) Since $0 < \alpha < \beta < 0.5$, this implies that we can in fact do much better than $O(\sqrt{N})$ buffer sizing [2] while maintaining all the good performance measures (full link utilization, negligible delay, and low loss).

C. Deriving Fluid Models Under Aggressive Scales

Suppose now that one wants to apply the existing stability criterion in the literature to the system under the aggressive scale. For example, according to [10], [15], [18], the criterion for linear stability of the system with queue-based marking (obtained from the Nyquist criterion) becomes $O(N^{1-\alpha}) < \zeta$, where ζ is a system-independent constant. On the other hand, the stability of all rate-based marking systems assuming Poisson type of packet arrivals requires that the slope of the marking function at the equilibrium be bounded, which then translates into the condition that the buffer size be bounded, i.e., $B(N) < \eta$ or at least $O(1)$ scaling should be employed for AQM [16], [11]. Note that for any $\alpha \in (0, 0.5)$ and large N, none of these criteria is satisfied, and this leads to a conclusion that the system under the aggressive scale is always unstable, albeit all the good performance measures as numerically observed. This is an embarrassing situation since an 'unstable' system still possesses all the good performance measures. So, one may ask: "What went wrong?," "What causes this self-conflicting conclusion?"

The above illustrates that one has to be very careful in choosing a "right" fluid model (or approximation) of the system under consideration. In general, depending on the scale of the system or types of limiting regimes employed, one can derive different fluid models (or approximations) that effectively capture the dynamics of original stochastic systems and then discuss their stability. For example, under the linear scale $O(N)$ for AQM and the buffer size, one can readily obtain fluid models for the normalized arrival rate and queue-length, where the marking is based on the normalized queue-length and the stability here refers to the convergence of the normalized queue-length [10], [15], [20], [21]. Similarly, in [13], [19] the authors derived a limiting system dynamics by considering random packet markings under $O(N)$ scale (but ignoring random packet arrivals within each RTT). On the other hand, under $O(1)$ scale with constant buffer size, the resulting fluid models are usually of rate-based types, where the marking is based on the arrival rate (e.g., an incoming packet is marked with probability $(\lambda/C)^{B}$ where packets arrive randomly with an average rate λ and buffer size B) and the stability now refers to the convergence of the arrival rate, from which the steady-state queue-length distribution can be inferred [16], [11], [5]. Further, in [17], [22], [23], where all the system parameters are kept constant (not dependent on N), some appropriate fluid models were obtained by sending the 'weight' of the exponential averaging for the queue-length to zero (or making it very small) and by making the exponentially averaged queue-length almost constant (separation of time-scales).

In this regard, it is now clear what went wrong in the aforementioned arguments. Specifically, the stability criterion in [10] is not meant to be used for systems with $O(N^{\alpha})$ scale, as it was originally derived for a system with $O(N)$ scale. Similarly, those in [17], [16], [11] are valid only under $O(1)$ scale (or under a rate-based marking). *Certainly, blind application of the fluid model could lead to very strange results.*

For the system with $O(N^{\alpha})$ scale, however, it is still unclear how to derive a "right" fluid model. This scale regime was discussed in [5], where the authors considered "underload" and "overload" situations separately and claimed that the system would be stable for moderate values of N, but tends to be unstable when N is very large (say, 5000 or more). However, this seems to conflict with the observation made in [2] and also our theoretical and simulation results later on in this paper showing that the system performance gets better for larger N. (See more discussion on this in Section VI-B.) As an attempt to address this "grey area", in this paper, we instead construct a fully stochastic model by taking the packet-level dynamics into account and analyze its performance in the steady-state, thus bypassing the difficulty of deriving a "right" fluid model for the system under the proposed aggressive scale.

III. System Model

A. A Doubly Stochastic Approach

In this section, we develop our doubly-stochastic approach to capture the randomness both in packet arrival instants and random packet markings. Consider a single link shared by N persistent flows, whose window sizes evolve according to the AIMD rule. Let T be the round-trip-time delay (RTT) for all N flows. We define $W_i^N(k)$ to be the window size of flow i at the start of kth RTT, where the superscript N means that there are N flows with capacity NC. Let $w_{\max}(w_{\max} > \mathrm{CT}+1)$ be the maximum window size for all flows, i.e., $1 \leq W_i^N(k) \leq w_{\max}$ for all i and N. Then, each window size $W_i^N(k)$ evolves as follows:

$$W_i^N(k+1) = \begin{cases} \left(W_i^N(k)+1\right) \wedge w_{\max}, & \text{if no marking} \\ \left\lfloor W_i^N(k)/2 \right\rfloor \vee 1, & \text{otherwise} \end{cases}$$

[3] Here, the steady-state should be interpreted in a distributional sense, not as a constant queue-length.

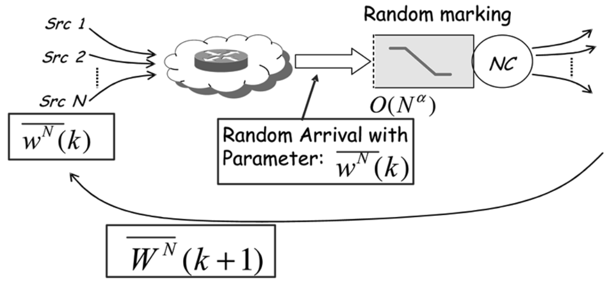


Fig. 2. Overview on our doubly stochastic approach.

where $x \wedge y := \min\{x, y\}$ and $x \vee y := \max\{x, y\}$. We define a set of window sizes for all N flows by $\overline{W^N}(k) := (W_1^N(k), W_2^N(k), \ldots, W_N^N(k))$.

Fig. 2 illustrates our doubly-stochastic approach. Suppose first that, at the start of the kth RTT, all the window sizes for N flows are given, i.e., $\overline{W^N}(k) = \overline{w^N}(k)$ where $\overline{w^N}(k) := (w_1^N(k), \ldots, w_N^N(k))$. (Note here that we use capital letters for random variables and lower case letters for their realizations.) Then each source i simply transmits $w_i^N(k)$ number of packets over each RTT onto the network at the source side. These packets traverse a number of links (upstream queues) being multiplexed with other flows, and suffer different amount of delay jitters. Thus, by the time they arrive to the queue of interest deep inside the network, their arrival times become random. In other words, *given* $\overline{W^N}(k) = \overline{w^N}(k)$, the arrival to the queue of interest becomes a random process with a set of parameters as a function of $\overline{w^N}(k)$, making the queue-length fluctuation also random.

Now, *for a given realization of the queue-length fluctuation*, we can find the distribution on whether incoming packets are marked or not from a specific marking function $p^N(x)$. Thus, by taking expectation over all possible queue-length realizations, we can calculate the distribution on whether there is any marked packet for flow i, which in turn determines the distribution of the window size for $(k+1)$th RTT duration, i.e., $W_i^N(k+1)$. Therefore, *given* $\overline{W^N}(k) = \overline{w^N}(k)$, it is possible to find the distribution of $\overline{W^N}(k+1)$ by capturing all the dynamics mentioned above, and as a consequence, there exists a Markovian structure between $\overline{W^N}(k)$ and $\overline{W^N}(k+1)$.

B. Model Description

In this section, we describe our model in detail and construct a Markov chain for $\overline{W^N}(k)$. As before, at the start of the kth RTT, suppose that all the window sizes are known. Then, in order to capture the random nature in packet arrivals, *given* $W_i^N(k) = w_i^N(k)$ for $i = 1, 2, \ldots, N$, we assume that the *actual* aggregate packet arrivals to the queue within one RTT can be modeled by a Poisson process with mean rate $\lambda_N(k)$ given by

$$\lambda_N(k) := \frac{1}{T} \sum_{i=1}^{N} w_i^N(k). \tag{1}$$

See Fig. 3 for illustration. Thus, the arrival process to the queue of interest is modeled by a *conditional Poisson process* (or *doubly stochastic* Poisson process) [24], [25]. We note that, under $\overline{W^N}(k) = \overline{w^N}(k)$, the Poisson assumption for the aggregate arrivals to the queue within one RTT is reasonable as long as packet arrivals from each flow are jittered independently at different upstream queues, due to the fact that the superposition of independent point processes under a suitable scaling converges weakly to a Poisson process [26], [27]. These Poisson packet arrivals over small time scale (sub-second scale or RTT-level) have also been empirically observed in [28]–[31]

Remark: Note that a conditional Poisson process is not a Poisson process since the rate over one RTT is given by the sum of the random window sizes. In other words, the rate itself is a random process, making the arrival process doubly-stochastic. Similar doubly-stochastic approaches for TCP/AQM systems have been used in [17], [11], [8]. However, results in [17], [11] are not applicable to systems under the aggressive scale, and in [8], the conditional Poisson process is used only to show that the process is dominated by a standard Poisson process with some

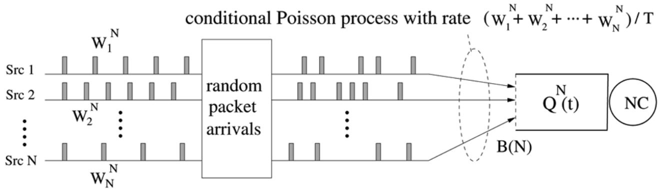

conditional Poisson process with rate $(W_1^N + W_2^N + \cdots + W_N^N)/T$
Src 1
Src 2
Src N
W_1^N
W_2^N
W_N^N
random packet arrivals
$Q^N(t)$
NC
B(N)

Fig. 3. Model for the arrival process to the queue within each RTT.

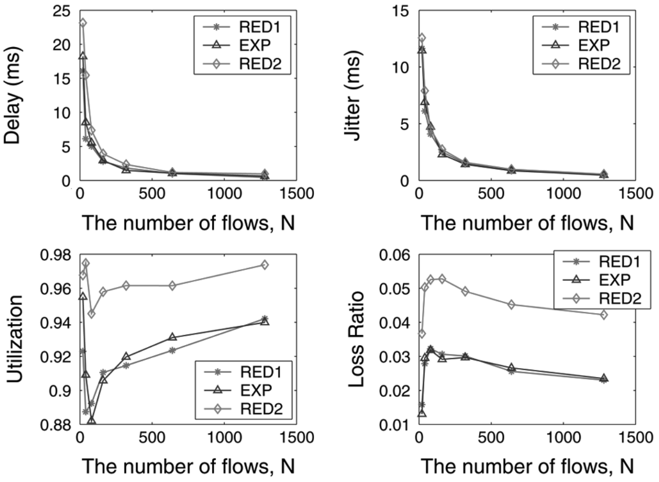


Fig. 4. Heterogeneous RTTs: Performance metrics with increasing number of flows (N) for fixed $\alpha = 0.2$.

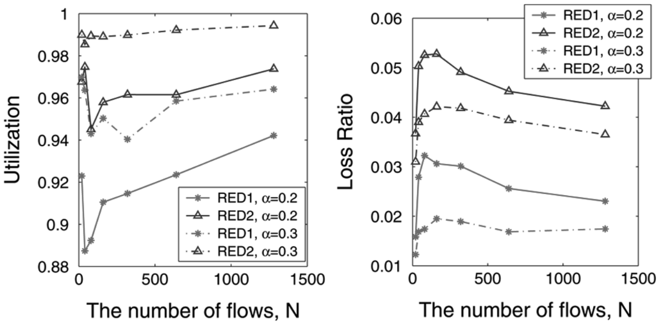


Fig. 5. Heterogeneous RTTs: Performance comparison of RED1 and RED2 with increasing number of flows (N) for different $\alpha (\alpha = 0.2$ and $0.3)$.

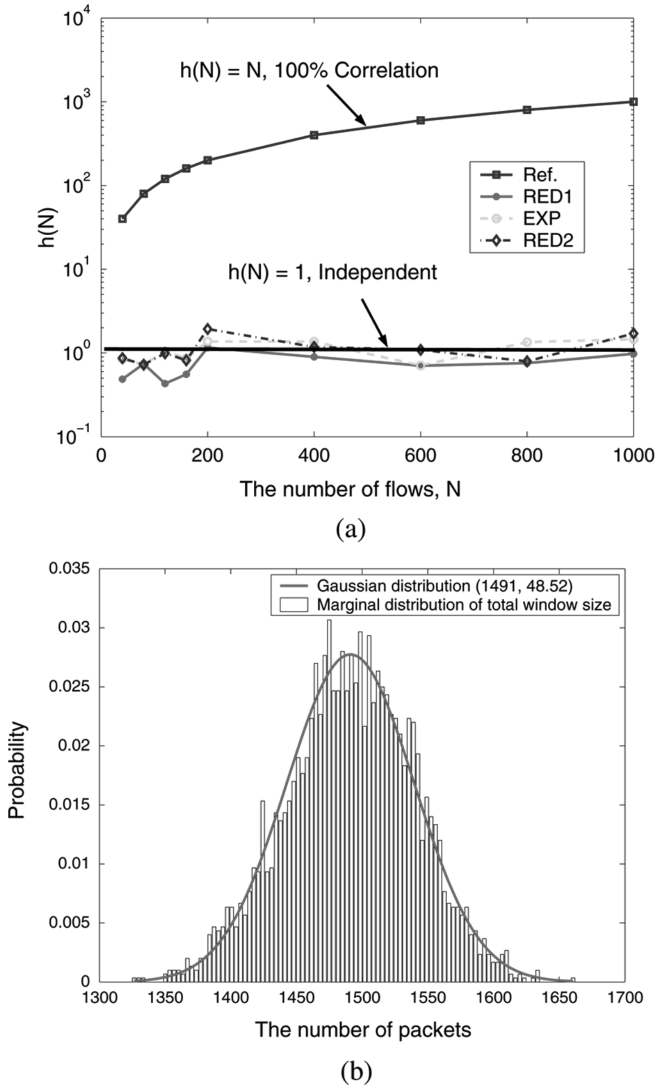


Fig. 6. (a) $h(N) = \mathrm{Var}\{\sum_{i=1}^{N} W_i^N\} / \sum_{i=1}^{N} \mathrm{Var}\{W_i^N\}$ as a function of N for different AQM schemes under $\alpha = 0.2$; (b) Histogram for aggregate window sizes from N flows under RED1, $N = 200$, and $\alpha = 0.2$. (a) $h(N)$. (b) PDF of aggregate window sizes.

larger constant rate (in terms of the maximum window size of each flow) and all the analysis was conducted via this standard Poisson process for the system under $O(1)$ scale with constant buffer size. Instead, we focus on different scalings and attempt to solve the constructed Markov chain for the doubly-stochastic model (random arrivals and random packet markings) in the steady-state with stationary distribution and derive the asymptotic performance results.

As before, $p^N(x)$ here is the probability that a packet is marked when the queue-length is x. In particular, our marking function $p^N(x)$ of scale $\alpha \in (0, 1/2)$ will satisfy the following: there exists a non-decreasing function $p : \mathbb{R}^+ \to [0, 1]$ such that, for all N and x

$$p^N(N^\alpha x) = p(x) \tag{2}$$

where

$$\lim_{x \to 0} p(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} p(x) = 1. \tag{3}$$

Note that this assumption is very general and the base marking function $p(x)$ can be any non-decreasing function with $p(0) = 0$ and $p(\infty) = 1$. See Fig. 1 for examples satisfying the above assumptions.

Within one RTT and *given* $\overline{W^N}(k) = \overline{w^N}(k)$, since we assume that the arrival is a Poisson process with mean rate $\lambda^N(k)$, we can find the distribution of the queue-length during the kth RTT by using any standard queueing model such as $M/D/1$ or $M/G/1$ as long as $\lambda_N(k) <$ NC. Further, we assume that if $\lambda_N(k) \geq NC$, all the incoming packets are marked.[4] Now, given $\overline{W^N}(k) = \overline{w^N}(k)$ during kth RTT, let $Q_{\overline{w^N}(k)}$ be the random variable denoting the queue-length fluctuation due to the random arrival instants. From our formulation above, we see that $Q_{\overline{w^N}(k)}$ is distributed according to the steady-state queue-length distribution of the $M/D/1$ or $M/G/1$ if $\lambda_N(k) < NC$ where $\lambda_N(k)$ is from (1). When $\lambda_N(k) \geq NC$, we can conveniently set $Q_{\overline{w^N}(k)} = \infty$ since $p(\infty) = 1$ from (3) and all the incoming packets will be marked. Also, this implies that $\sum_{i=1}^{N} W_i^N(k) < NCT + N$ all the time, since all the flows will drop their rates by half once the sum of the window sizes goes beyond NCT.

Now, *given* $\overline{W^N}(k) = \overline{w^N}(k)$ and *given* a realization of queue-length, each packet will be marked with probability $p^N(Q_{\overline{w^N}(k)})$. Since flow i transmits $w_i^N(k)$ packets, we see that flow i receives no mark during the kth RTT with probability

$$p_{i,N,k}\mathbb{E} := {}_{-}Q \left\{ \left[1 - p^N \left(Q_{\overline{w^N}(k)} \right) \right]^{w_i^N(k)} \right\} \tag{4}$$

where $\mathbb{E}_Q\{\cdot\}$ is the expectation over all possible realizations of the queue-length under $\overline{W^N}(k) = \overline{w^N}(k)$.

Given $\overline{W^N}(k) = \overline{w^N}(k)$, suppose first $\lambda_N(k) <$ NC. Then the queue remains stable within that RTT (i.e., utilization < 1), and each packet will be marked independently of any other. We thus expect that the events of flow i receiving at least one marked packet are likely to be independent for different i, which implies independence of window sizes among different flows at the next RTT.

In the case of $\lambda_N(k) \geq$ NC, since all the flows will back off in the next RTT, the window sizes at the next RTT will not be independent. However, we will show that this event becomes rare for large N in the sense that $\mathbb{P}\{\sum_{i=1}^{N} W_i^N(k) \geq NCT\} \to 0$ as N increases (see Lemma 1). Hence, for any case, we can assume the following:

Assumption 1: Given $\overline{W^N}(k) = \overline{w^N}(k)$, the random variables $W_i^N(k+1) (i = 1, 2, \ldots, N)$ are independent.

Indeed, our simulation results in Section V.C will also show that the window sizes for different flows are mostly independent even under our aggressive marking scale (see Fig. 6). Further, Assumption 1 will be used only to construct a Markov chain for the system and will never be used again once the system enters a steady-state.

Now, for any given N, $\{\overline{W^N}(k)\}_{k \geq 0}$ forms an N-dimensional homogeneous Markov chain with its state space given by

$$E := \{1, 2, \ldots, w_{\max}\}^N \cap S(N) \tag{5}$$

[4] This corresponds to an unstable queue with utilization $\rho \geq 1$. Thus, for large N, the queue-length will explode for any marking scale N^α $(0 < \alpha < 1/2)$ within that RTT and almost all packets will be marked.

where

$$S(N) = \left\{ \overline{W^N}(k) | N \leq \sum_{i=1}^{N} W_i^N(k) < N(CT+1) \right\}. \tag{6}$$

From Assumption 1, the transition probability is given by

$$\begin{aligned} \mathbb{P}\{\overline{W^N}(k+1) =& \overline{w^N}(k+1) | \overline{W^N}(k) = \overline{w^N(k)}\} \\ &= \prod_{i=1}^{N} \text{pr} \left\{ W_i^N(k+1) = w_i^N(k+1) | \overline{W^N}(k) \right. \\ &\qquad \left. = \overline{w^N}(k) \right\} \end{aligned} \tag{7}$$

where

$$\begin{aligned} \mathbb{R} &\left\{ W_i^N(k+1) = w_i^N(k+1) | \overline{W^N}(k) = \overline{w^N}(k) \right\} \\ &= \begin{cases} p_{i,N,k}, & \text{if } w_i^N(k+1) = \left(w_i^N(k)+1\right) \wedge w_{\max} \\ 1 - p_{i,N,k}, & \text{if } w_i^N(k+1) = \left\lfloor w_i^N(k)/2 \right\rfloor \vee \\ 10, & \text{otherwise} \end{cases} \end{aligned} \tag{8}$$

and $p_{i,N,k}$ is from (4).

In the next section, we will investigate the limiting behaviors of this Markov chain in the steady-state as N increases and provide key performance results under the aggressive scale.

IV. Theoretical Results

From the Markov chain defined in (7) and (8) with its state space given by (5) and (6), it is not difficult to see that the chain is irreducible and aperiodic. Further, for any given N, the state space E is finite. Thus, from Theorem 3.3 in [32, p. 105], the chain is positive recurrent, hence is ergodic. So, there exists a unique stationary distribution π, and the process $\{\overline{W^N}(k)\}_{k \geq 0}$ converges *in variation* to π [32], [33]. In other words, we have

$$\lim_{k \to \infty} \sum_{w^N \in E} \left| \mathbb{P}\left\{ \overline{W^N}(k) = \overline{w^N} \right\} - \pi \left\{ \overline{w^N} \right\} \right| = 0. \tag{9}$$

This proves that, regardless of initial distributions of window sizes of N flows, the chain converges to a steady-state in which the process $\overline{W^N}(k)$ has a stationary distribution π. In the steady-state, we basically have a steady-state queueing system, where the input process to the queue is a *conditional* Poisson process (which has stationary increments, but not independent increments) and its underlying process $\overline{W^N}(k)$ is a homogeneous, ergodic Markov chain with a stationary distribution π. As the distribution of $\overline{W^N}(k)$ does not depend on k in the steady-state, we will use $\overline{W^N}$ to denote the window sizes random variables in the steady-state and π to denote the distribution for $\overline{W^N}$ whenever there is no ambiguity.

In principle, since we know the transition matrix from (7) and (8), it is theoretically possible to find out the stationary distribution π by solving a set of balance equations. However, due to the complicated structure for the transition matrix (especially for large N), direct calculation of the steady-state distribution is computationally prohibitive. Instead, our approach here is to derive some properties of key performance metrics of the system as N increases, without explicitly solving for the stationary distribution.

From (9), since $(1-p^N(\cdot))^{W_i^N}$ is bounded, we know that the (unconditional) probability of flow i not marked during the kth RTT also converges to the following steady-state probability of flow i not marked:

$$F_i^N := \mathbb{E}\left\{ \left[1 - p^N(Q_{\overline{W^N}})\right]^{W_i^N} \right\} \tag{10}$$

where the expectation is taken both over $\overline{W^N}$ and random queue-length fluctuation. Specifically, we can calculate (10) as follows:

$$F_i^N = \mathbb{E}_{\overline{W^N}} \left\{ \mathbb{E}_Q \left\{ \left[1 - p^N(Q_{\overline{W^N}})\right]^{W_i^N} \middle| \overline{W^N} \right\} \right\}.$$

The following result will be used in proving our main results. Due to space limitation, we only provide an outline of the proof in this paper. Please see our technical report [34] for complete proofs.

Proposition 1: For any given $N > 0$ and i $(1 \leq i \leq N)$, we have

$$\frac{\mathbb{E}\left\{W_i^N\right\} - 1}{w_{\max} + 1} \leq \mathbb{E}\left\{ \left[1 - p^N(Q_{\overline{W^N}})\right]^{W_i^N} \right\}. \tag{11}$$

Further, there exists a constant $B \in (0,1)$ (independent of N) such that

$$\mathbb{E}\left\{ \left[1 - p^N(Q_{\overline{W^N}})\right]^{w_{\max}} \right\} \leq B < 1, \quad \forall N > 0. \tag{12}$$

Sketch of the Proof: We basically want to show that the expression in (10) should not be too close to 0, nor to 1. To see this, suppose that $F_i^N \approx 0$ for some large N. Then, for most sample paths (realizations), each flow i will have very high probability of being marked and so most of N flows will decrease their window sizes by half. Thus, the distribution of the window sizes for the next RTT duration will be different from the current one, and this contradicts that the system is in steady-state with a stationary distribution. Similar arguments hold for the case if it is too close to 1.

We define the steady-state utilization of the system by

$$\rho_N = \rho(N) := \frac{1}{\text{NCT}} \mathbb{E}\left\{W_1^N + W_2^N + \cdots + W_N^N\right\}. \tag{13}$$

As the system is in the steady-state with a stationary arrival process to the queue (a conditional Poisson process), we should have $\rho(N) < 1$ for all N, since otherwise the queue-length will increase without bound almost surely and eventually *all* the flows will drop their rates by half, contradicting the fact that the distribution of window sizes does not depend on time.

In the steady-state, we already observed that the stationary arrival process to the queue of interest becomes the conditional (or doubly stochastic) Poisson process where the underlying process is a stationary, ergodic Markov chain. Hence, *given* $\overline{W^N}(k) = \overline{w^N}(k)$, within that RTT duration, the arrival process to the queue becomes a Poisson process with rate $\lambda_N(k)$ as in (1). In general, *under* $\overline{W^N}(k) = \overline{w^N}(k)$, this Poisson characterization within that RTT enables us to describe the queueing dynamics in terms of $\lambda_N(k)$ and any general service time distribution G. However, in order to prevent the analysis from being intractable and unwieldy, we will assume that the service time distributions of each packet are independent and identically distributed (i.i.d.) and exponentially distributed throughout the rest of the paper. In this way, given $\overline{W^N}(k) = \overline{w^N}(k)$, the queue dynamics can be described by that of $M/M/1$ queue within each RTT.[5]

Let us define $X_i^N = W_i^N/\rho(N)CT$ such that $\mathbb{E}\{X_i^N\} = 1$, and also their scaled sum Y_N by

$$Y_N := \frac{X_1^N + X_2^N + \cdots + X_N^N - N}{\sqrt{N}}. \tag{14}$$

We will then assume the following:

Assumption 1: The distribution of Y_N for large N is well-behaved in the sense that for any positive sequence θ_N with $\theta_N \downarrow 0$ as $N \uparrow \infty$, we have

$$\limsup_{N\to\infty} \mathbb{E}\{e^{\theta_N Y_N}\} < \infty. \tag{15}$$

Remark 2: Assumption 2 is quite technical and requires the moment generating function (with vanishing exponent θ_N) be simply finite. It basically states that the scaled sum of W_i^N does not quickly diverge as N grows. Indeed, if W_i^N for $i = 1, 2, \ldots, N$ are i.i.d. or weakly-dependent, then by the CLT, for large N, the distribution of Y_N will not be so much different from a Gaussian random variable, for which (15) trivially holds. Moreover, this assumption is readily satisfied in practice as we will show in Section V-C that the window sizes for different flows behave as if they are more or less independent under our aggressive marking scale (see Fig. 6).

Our first main result shows that the utilization of the steady-state system converges to 1 as N increases.

Theorem 1: Under Assumption 2, we have

$$\lim_{N\to\infty} \rho(N) = 1.$$

Sketch of the Proof: Since $\rho(N) < 1$ for each N, we only have to show that $\liminf_{N\to\infty} \rho(N) = 1$. From (12) and the convexity of $(1-x)^{w_{\max}}$ for $x \in [0,1]$, we can show by using Jensen's inequality that the packet marking probability is lower-bounded. Specifically, there exists a constant $A \in (0,1)$ such that $A \le \mathbb{E}\{p^N(Q_{\overline{W^N}})\}$. Since $p^N(N^\alpha x) = p(x)$ is non-decreasing in x, we have $p(x) \le p(q) + (1-p(q))1_{\{x>q\}}$ for all positive x, q. Then, by choosing $x = Q_{\overline{W^N}}/N^\alpha$ and taking expectation, we can similarly show that there exists a constant B such that $0 < B \le \mathbb{P}\{Q_{\overline{W^N}} \ge qN^\alpha\}$ for some $q \in (0,\infty)$.

Since the queue dynamics is modeled by $M/M/1$ under $\overline{W^N} = \overline{w^N}$, we can show by using the conditional expectation that

$$P\{Q_{\overline{W^N}} \ge n\} = \mathbb{E}\left\{\left(\frac{W_1^N + W_2^N + \cdots + W_N^N}{NCT}\right)^n \wedge 1\right\} \tag{16}$$

for any $n \ge 0$. Set $n = qN^\alpha$ and note that the above is bounded by

$$\mathbb{E}\left\{\left(\frac{W_1^N + W_2^N + \cdots + W_N^N}{NCT}\right)^{qN^\alpha}\right\}. \tag{17}$$

Now, observe that (17) can be rewritten as

$$(\rho(N))^{qN^\alpha} \times \mathbb{E}\left\{\left(1 + \frac{Y_N}{\sqrt{N}}\right)^{qN^\alpha}\right\}. \tag{18}$$

Assumption 2 guarantees that the second term in (18) is finite. In other words, there exists a constant D such that $0 < D \le (\rho(N))^{qN^\alpha}$ for all sufficiently large N. In other words, by taking logs, we must have

$$-\infty < \liminf_{N\to\infty} N^\alpha \log \rho(N) \tag{19}$$

which readily implies $\liminf_{N\to\infty} \rho(N) = 1$. This completes the proof. ■

Theorem 1 guarantees that $\mathbb{E}\{W_i^N\} \approx C > 1$ for all large N. Since $1 \le W_i^N \le w_{\max}$ and from Proposition 1, it is straightforward to see that there exist constants a, b such that

$$0 < a < \mathbb{E}\left\{p^N(Q_{\overline{W^N}})\right\} < b < 1, \quad \text{for all large } N \tag{20}$$

(also, see proofs in [34]). In other words, the expected marking probability of each packet is bounded away from 0 and 1 uniformly over N.

Before we proceed to our second main result, we present the following lemma.

Lemma 1: Suppose $CT > 3$. Then, under Assumption 2, we have

$$\lim_{N\to\infty} \mathbb{P}\left\{\sum_{i=1}^{N} W_i^N \ge NCT\right\} = 0. \tag{21}$$

Sketch of the Proof: First, since the Markov chain is ergodic, for any bounded function f, we have from Theorem 4.1 in [32, p.111] that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(\overline{W^N}(k)) = \mathbb{E}_\pi\left\{f(\overline{W^N})\right\} \text{ a.s.} \tag{22}$$

where $\mathbb{E}x_\pi$ is the expectation with respect to the stationary distribution (π) of $\overline{W^N}$. Suppose that (21) is not true. Then, with strict positive probability, the sum of the window sizes becomes at least NCT. Whenever this happens, we know that all the flows receive marks, so they all reduce their window sizes by half in the next RTT. Now, observe that once all of them are halved, it will take a certain number of RTTs for the sum of the window sizes to cross NCT again, as the sum of the window sizes can increase at best only by N per RTT. During that period (i.e., an interval between two consecutive "crossing the NCT" events), the average window size remains smaller than NCT, and this type of behavior will repeat forever. So, by taking time average and using (22), we arrive to the conclusion saying that the steady-state utilization is strictly bounded away from 1, which contradicts the result in Theorem 1 for large N. Therefore, Lemma 1 follows. See [34] for the complete proof. ■

[5] Given $\overline{W^N}(k) = \overline{w^N}(k)$, the queue dynamics can be similarly analyzed via $M/M/1/K$ or $M/D/1/K$ (with a buffer of size $O(N^\beta)$. In order to avoid unnecessary complicated derivations, however, we use $M/M/1$ as an approximation. Note also that the loss probability of a finite buffer of size K is bounded by the overflow probability of an infinite-buffer with level K.

We now present our second main result, which states that the fluctuation of the steady-state queue-length is no more than $O(N^{\alpha+\epsilon})$.

Theorem 2: Let $Q_{\overline{W^N}}$ be the steady-state queue-length random variable. Suppose that $CT > 3$ and Assumption 2 is satisfied. Then, for any $\epsilon > 0$, we have

$$\lim_{N\to\infty} \frac{Q_{\overline{W^N}}}{N^{\alpha+\epsilon}} = 0 \quad \text{in probability.} \tag{23}$$

Sketch of the Proof: Similarly as in the Proof of Theorem 1, we can show from Proposition 1 that there exists a constant η such that $\mathbb{P}\{Q_{\overline{W^N}} \geq q'N^\alpha\} \leq \eta > 1$ for some $q' \in (0,\infty)$. Then, from (16) with $n = q'N^\alpha$ and from Lemma 1, it is not difficult to see that

$$\mathbb{E}\left\{\left(\frac{W_1^N + W_2^N + \cdots + W_N^N}{NCT}\right)^{q'N^\alpha}\right\} \leq \eta' < 1$$

for some constant η' and for all sufficiently large N. Using the same steps in the last part of the Proof of Theorem 1, we obtain

$$\limsup_{N\to\infty} N^\alpha \log\rho(N) < 0. \tag{24}$$

Further, since $0 < \alpha < 1/2$, without loss of generality, we only have to prove (23) for $\epsilon > 0$ with $\alpha + \epsilon < 1/2$. Note that, from (24), there exists $\kappa > 0$ such that $N^\alpha \log\rho(N) \leq -\kappa$ for all sufficiently large N. Thus

$$N^{\alpha+\epsilon}\log\rho(N) \leq -\kappa N^\epsilon, \quad \text{for all sufficiently large } N. \tag{25}$$

Now, as before, from (16) with the choice of $n = \delta N^{\alpha+\epsilon}$, it follows that $\mathbb{P}\{Q_{\overline{W^N}} > \delta N^{\alpha+\epsilon}\}$ is bounded by

$$\mathbb{E}\left\{\left(\frac{\sum_{i=1}^N W_i^N}{NCT}\right)^{\delta N^{\alpha+\epsilon}}\right\} + \mathbb{E}\left\{\sum_{i=1}^N W_i^N \geq NCT\right\}. \tag{26}$$

From Assumption 2 and (25) and by the argument used in (18), the first term in (26) can be bounded by $M \cdot \exp(-\delta\kappa N^\epsilon)$ for some constant $M < \infty$. Further, Lemma 1 asserts that the second term in (26) also decreases to zero, and we are done. See [34] for the complete proof. ■

Remarks 3: Relations (19) and (24) tell us how fast $\rho(N)$ converges to 1. Specifically, we can rewrite (19) and (24) as $\log\rho(N) = O(1/N^\alpha)$, or $\rho(N) \approx \exp(-\kappa N^{-\alpha})$ for some constant $0 < \kappa < \infty$. Note that the speed of convergence depends on the scaling parameter α. In particular, it can be quite slow for very small α.

In Theorem 2, the positive constant ϵ plays a critical role to ensure that $Q_{\overline{W^N}}$ is no larger than $O(N^{\alpha+\epsilon})$. From Proposition 1 and Theorem 1, we can see that $\mathbb{E}\{p^N(Q_{\overline{W^N}})\}$ is always bounded away from 0 and 1 (uniformly over N). Since the marking function is scaled on the order of N^α (Recall $p^N(N^\alpha x) = p(x)$), this in turns implies that $\mathbb{P}\{Q_{\overline{W^N}} \in O(N^\alpha)\}$ is also bounded away from 0 and 1 uniformly over N. Thus, if we set the buffer size to $O(N^\alpha)$, the loss probability will not go to zero. Instead, the added "margin" of length $O(N^\epsilon)$ is necessary to make the loss probability go to zero under the buffer of size $O(N^{\alpha+\epsilon})$ and to ensure that all the "congestion signals" are from packet marking, not from packet loss.

TABLE I
AQM Parameters for *ns*-2 Simulations

AQM	Parameters
RED1	$q_{min}N^\alpha = 2N^\alpha$, $q_{max}N^\alpha = 10N^\alpha$ $P_{max} = 0.2$, Buffer Size $B(N) = 12N^{\alpha+\epsilon}$
EXP	$\gamma = -10/\ln(1-P_{max})$, Buffer Size = $12N^{\alpha+\epsilon}$
RED2	$q_{min}N^\alpha = 4N^\alpha$, $q_{max}N^\alpha = 11N^\alpha$ $P_{max} = 0.05$, Buffer Size $B(N) = 12N^{\alpha+\epsilon}$

Remark: From the practical point of view, however, the added margin $O(N^\epsilon)$ may grow very slowly. For example, if $\epsilon = 0.1$, then even for $N = 1000$ flows, we have $N^\epsilon \approx 2$, and this margin (a factor of 2) may not be large enough to ensure almost zero packet loss. In other words, in Theorem 1, the speed of convergence to zero can be quite slow. But, we emphasize that our results provide theoretical basis and the loss probability can be made indeed arbitrarily small for any given constant $\epsilon > 0$ and for all "sufficiently large" N.

V. Numerical Results

In this section, we provide numerical results using ns-2 [35] to validate our results in Section IV-C. First, we show that even under our aggressive scale, the window sizes of each of N flows tend to be independent or weakly correlated, which is required for our theoretical analysis. Second, we show that as the number of flows and the size of the link capacity increase,[6] the queueing delay becomes smaller, the link utilization increases, and the packet loss probability decreases, although the convergence for the loss probability is a bit slow as predicted.

Throughout the simulations, we consider the following four performance metrics: 1) average queueing delay; 2) delay jitter (we calculate the standard deviation from all the measured queueing delays); 3) link utilization; and 4) packet loss ratio.

A. AQM Configurations

We consider three different queue-based AQM schemes with $O(N^\alpha)$ scales for our simulations: two RED schemes under different configurations (referred by RED1 and RED2) and EXP as in Fig. 1(b) . We set all the thresholds in the queue-length to be proportional to N^α, (i.e., $p^N(N^\alpha x) = p(x)$). Table I summarizes the parameters of each AQM scheme used in our simulation.

For RED1, the minimum and the maximum thresholds ($q_{\min}N^\alpha$ and $q_{\max}N^\alpha$ in Fig. 1(a)) are $2 \times N^\alpha$ and $10 \times N^\alpha$, respectively. The maximum dropping probability, $P_{\max}$, is set to 0.2 and the queue averaging factor, q_weight_, is set to 1. (We have also run simulations with other queue averaging factors, e.g., q_weight_ = 0.002, and obtained similar results.) As to EXP, we use a queue-based AQM with an exponential marking profile as in Fig. 1(b)). In order to make a fair comparison with RED, we choose $\gamma = -\frac{10}{\ln(1-P_{\max})}$ (as in Fig. 1(b) such that $p^N(\max_{\text{th}}) = p^N(10N^\alpha) = P_{\max}$. RED2 is similar to RED1, but with different thresholds and $P_{\max}$. In any case, the buffer

[6] Note that each flow will get approximately the same bandwidth share regardless of the size of the system N.

size in our simulations is set to $B(N) = 12 \times N^{\alpha+\epsilon}$ with $\epsilon = 0.1$, as the margin $O(N^{\epsilon})$ is required in Theorem 1.

B. Simple Dumbbell Topology

We first consider a simple dumbbell topology with a single bottleneck link shared by N long-lived TCP flows. We set the sum of two-way propagation delays for N flows to be i.i.d. and uniformly distributed over [120, 280] ms (with mean 200 ms, as in [2]). Each packet size is fixed to 500 bytes and the link capacity is $NC = N \times 100$ kbps. Thus, the target throughput for *each flow* is $C = 100$ kbps or 25 packets/sec.

Fig. 4 shows the four performance metrics as the number of long-lived flows (N) increases under our aggressive scales for AQM schemes with $\alpha = 0.2$ and $\epsilon = 0.1$.[7] Note that for all the schemes considered, the queueing delay and delay-jitter decrease sharply with the increase of N, and the utilization and the packet loss ratio show clear trend with the number of flows N, as predicted by Theorems 1 and 2. In addition, RED1 and EXP yields almost identical performance in all performance metrics, while we observe some tradeoffs between RED1 (or EXP) and RED2. RED1 and EXP have lower packet-loss ratio with the smaller link utilization, while RED2 produces higher link utilization with a little larger packet-loss ratio.

Fig. 5 shows the link utilization and the loss ratio for RED1 and RED2 under $\alpha = 0.2$ and $\alpha = 0.3$. We see that for larger α, the system yields higher link utilization and smaller loss ratio. This is reasonable because larger α implies that the corresponding buffer size and the margin for packet marking $O(N^{\alpha})$ is larger. In general, for any given N, it is possible to improve the link utilization at the expense of higher loss ratio (or vice versa) by using different AQM configurations. But, under $O(N^{\alpha})$ scale, all these differences will disappear when N becomes larger, since the link utilization will approach to 1 and the loss ratio to zero as shown in Section IV.

C. Independence of Window Sizes Among Different Flows

In Sections III and V, we have assumed that there exists independence (or weak dependence) of window sizes random variables among different flows. Those assumptions, in conjunction with random packet arrivals within each RTT, play a crucial role in establishing our main results in Section IV. In this section, we show that the window sizes are indeed almost independent under all AQM schemes with our aggressive scaling.

To investigate whether there exists any strong correlations among $W_i^N, i = 1, 2, \ldots, N$, we consider a function $h(N)$ defined by

$$h(N) := \frac{\mathrm{var}\left\{\sum_{i=1}^{N} W_i^N\right\}}{\sum_{i=1}^{N} \mathrm{var}\left\{W_i^N\right\}}.$$

If the window sizes are independent for all N, we immediately have $h(N) = 1$ for all N. If they are all "perfectly" correlated (synchronized), we would have $h(N) \approx N$.

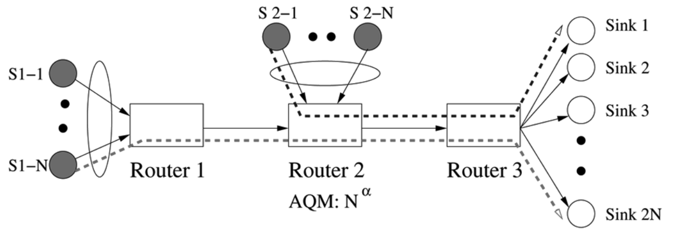


Fig. 7. Simulation topology: multiple bottleneck links with heterogeneous RTTs.

From the ns-2 simulation results in Section V-B, we measured the values $h(N)$ as N varies for all the AQM schemes considered, and plot them in Fig. 6(a) on a semi-log scale.[8] Fig. 6(b) shows the corresponding histogram for aggregate window sizes from all N flows under RED1 and $\alpha = 0.2$. As we see, the function $h(N)$ mostly stays around 1 for all AQM schemes considered over all ranges of N. This implies that there exists virtually no positive correlation among the window sizes for different flows even under the aggressive marking $(\alpha = 0.2)$. From Fig. 6, the pdf of the total window size is shown to be very close to a Gaussian distribution with mean 1419 and standard deviation 48.52 packets. We also measure the function $h(N)$ and the histogram under various configurations with N up to 1000 flows and different $\alpha \in (0, 0.5)$. In all these cases, we observe that the function $h(N)$ is always around 1 and the histogram closely matches with some Gaussian marginal distribution, suggesting that the window sizes for all flows are more or less independent all the time.

D. Multiple Links With Heterogeneous RTTs

Up until now, we have simulated various network scenarios with a single link. In this section, we present simulations results using multiple links in the network.

Fig. 7 depicts a network topology with multiple links of our consideration. In this scenario, we have three routers and two groups of flows, where group 1 (S1–1 to S1–N) traverses routers 1–3 and group 2 (S2–1 to S2–N) goes through routers 2 and 3. We use drop-tail for routers 1 and 3 with $O(N)$ buffer sizes, and for router 2, we use different AQM schemes as before with the scale of N^{α}. The capacities of routers 1 and 2 are $N \times 100$ and $2N \times 100$ kbps, respectively. The two-way propagation delays of group 1 flows are uniformly distributed over [110, 190] ms with mean 150 ms, and over [60, 140] with mean 100 ms for group 2.

As before, Fig. 8 shows the similar trends for all the performance metrics measured at router 2, which demonstrate that our results hold for more general network configurations. Regarding the link utilization and packet loss ratio, we see some zigzags when the number of flows is small (up to around 200). For such small values of N, it is still far away from the domain of convergence for Theorems 1 and 2, and the performance is mainly governed by other factors. For example, the buffer size in Table I is $12 \times N^{\alpha+\epsilon}$, and if N is not large enough, the resulting queue

[7] In our numerical results, we were unable to simulate the case of N larger than 1300 flows mainly because of the limitation in the ns-2.

[8] We sampled the window size of each flow i every 20 ms except for some initial period. Since RTTs are between 120 and 280 ms in our simulation setup, we get 6 to 14 samples of the window size of each flow per RTT.

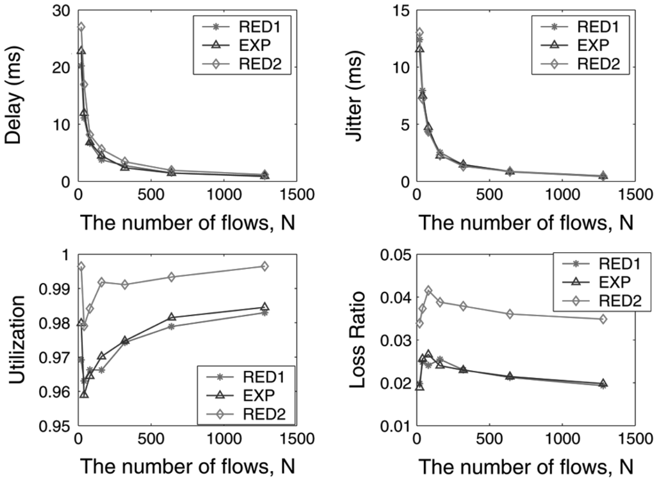


Fig. 8. Multi-hop topology: Performance metrics with increasing number of flows (N) for fixed $\alpha = 0.2$ and $\epsilon = 0.1$.

size will be more affected by the preceding factor, 12, rather than by $N^{\alpha+\epsilon}$.

VI. Discussion

A. Importance of Random Arrivals for Aggressive Scale

In Section II we already noticed that different ways of scaling lead to different modeling of the system. In particular, random packet arrivals over sub-RTT levels play a crucial role in capturing the dynamics of systems with $O(1)$ scale [16], [11], [5], while coarser time-scale models capturing only RTT-level dynamics suffice for systems with $O(N)$ scale [10], [15], [13], [19]. In this section, we further illustrate that the random packet arrivals, in conjunction with the random packet marking, are indispensable components in our modeling for systems under the aggressive scale.

Suppose that we were to ignore the randomness in packet arrivals within each RTT and consider only the effect of random packet marking in our modeling. As observed in Section V-C, under the aggressive scale, the window size random variables for N flows are mostly independent and the aggregate window size is best matched by a Gaussian distribution with standard deviation of $O(\sqrt{N})$. While the buffer size and the region for marking are smaller than $O(\sqrt{N})$, one may think that, by appropriately 'shifting' the Gaussian marginal distribution, it might be still possible to explain the high link utilization and low packet loss via the CLT type of argument in the steady-state. In this case, the average marking probability of a packet can be calculated by integrating the Gaussian marginal distribution over a region where the marking takes place, i.e., an interval I of length $O(N^{\alpha})$ (See Fig. 9). Thus, as $N \rightarrow \infty$, the average marking probability will become

$$\begin{aligned}\{\text{marking}\} &= \int_I \frac{1}{\sqrt{2\pi O(\sqrt{N})}} e^{-\frac{(x-N\mu)^2}{2O(N)}}\, dx \\ &\leq \frac{1}{\sqrt{2\pi O(\sqrt{N})}} \int_I 1\, dx \\ &= \frac{O(N^{\alpha})}{\sqrt{2\pi O(\sqrt{N})}} \longrightarrow 0 \qquad (27)\end{aligned}$$

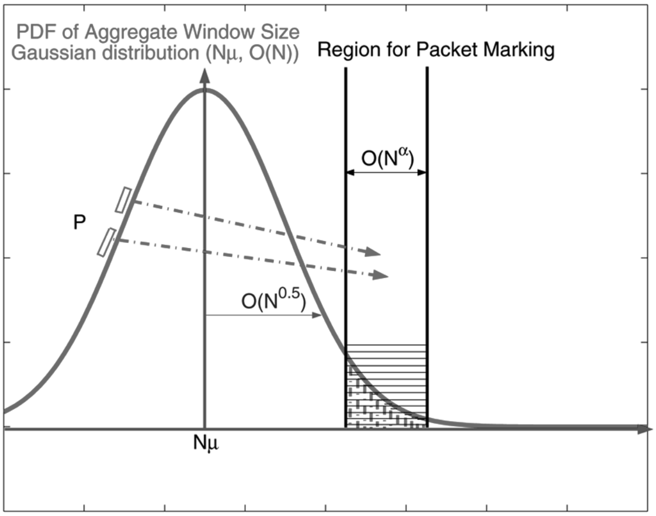

PDF of Aggregate Window Size
Gaussian distribution (Nμ, O(N))
Region for Packet Marking
O(N^α)
P
O(N^0.5)
Nμ

Fig. 9. If we consider only the Gaussian marginal distribution of the aggregate window sizes, packet marking probability goes to zero as the region for packet marking $O(N^{\alpha})$ is too small compared to $O(\sqrt{N})$. However, under the random packet arrivals, packets 'P' outside of the marking region may also be marked as the random arrivals may create temporary queue buildup.

since $\alpha < 1/2$. This means that the probability of packet marking goes to zero as N becomes larger, and, therefore, the system will never be in the "steady-state" as (27) clearly contradicts (20). Thus, independence among flows alone is not sufficient to explain the good performance of the TCP/AQM system under the aggressive marking scale. This is because the CLT-based argument lacks the dynamics of random packet arrivals within each RTT. For instance, in Fig. 9, consider packets "P" located on the left side of the distribution. While these packets are outside of the marking region and thus will never be marked under the CLT-based approach, they may be so under the random packet arrivals as these may create temporary queue buildup. In fact, due to the random arrival effects, we note that marking can take place *anywhere* on the x-axis in Fig. 9. So, in some sense, the integration in (27) should have been taken over a much larger interval (rather than over the interval I of length $O(N^{\alpha})$) such that the marking probability does not vanish as shown in (20).

B. Related Work on Hybrid Approach

In the literature, there have been several attempts to integrate stochastic components into the fluid modeling for TCP/AQM systems under various scales [11], [5]. In [11], the authors considered the usual fluid recursion driven by a Poisson input and obtained limiting deterministic equations of system dynamics for the window size evolutions as the system size increases. They considered $O(N)$ and $O(1)$ for packet marking and showed that the system behaves as if it were a rate-based one under $O(1)$ scale.

On the other hand, in [5], the authors considered all possible ways of scaling the buffer size, i.e., $B(N) = O(N^{\gamma})$ with $\gamma = 0, \gamma = 1$, and $0 < \gamma < 1$. The approach is still based on the fluid model, but with different reasoning obtained from the stochastic queueing theory for different γ. Specifically, for $0 < \gamma < 1$, the authors applied different *open-loop* queueing theories with differen traffic patterns for underload ($\rho < 1$) and overload ($\rho > 1$) cases, where ρ is the link utilization. Then, the authors proceeded to show that the loss probability decreases to zero for underload and the marking function is given by $p = [1 - 1/\rho]^+$ for overload. In other words, there is no meaningful congestion signal when $\rho < 1$.

However, this approach is somewhat questionable. As we have shown in this paper, the randomness in packet arrivals ensures that, even when $\rho < 1$, there always exist some packet marks or losses that contribute to giving some "feedback" to all the senders and the amount of feedback will be larger as the link utilization gets closer to one (but still less than one). We show that the link utilization actually increases to one under such a scale for the buffer size and AQM. Thus, we believe that the dichotomy between underload and overload without considering the randomness in packet arrivals may not be suitable for capturing the real dynamics of TCP/AQM system, especially when the system operating point is very close to the boundary ($\rho = 1$), and that a direct analysis of the doubly stochastic, closed-loop model for the system will do a better job. Moreover, it is argued in [5] that although the scaling of $0 < \gamma < 1$ would work fine when N is moderately large (say, a few hundreds, as is also shown in [2]), the system will become unstable with possibly poor performance when N is very large, say, 5000 or more.[9] While we have not been able to simulate such a large N ($N \geq 5000$), however, based on the trend in our simulation results in Section V and the results in [2], it is rather hard to believe that instability appears all of a sudden when N is larger than some threshold. Again, this observed disparity seems to corroborate the difficulty of deriving a right fluid model or representation of the system under the proposed aggressive scale.

VII. Conclusion

In this paper, we have investigated the performance of TCP/AQM systems with ECN marks under the aggressive scale for packet marking and a buffer of size smaller than $O(\sqrt{N})$. Under such a scale, we have demonstrated that randomness both in packet arrivals and in packet marking must be taken into account and developed a doubly stochastic model to capture all the system dynamics. Based on our model, we have proven that the system yields good performance in terms of full link utilization and almost zero packet loss, while making the queueing delay negligible and saving more buffer space, as the system size increases. We have also provided an extensive set of numerical results using ns-2 under a variety of network configurations with different AQM schemes, and further validated our results in more realistic settings. Interestingly enough, it turns out that our findings of all the good performance somewhat contradict to a recent result implying that the system under the proposed scale will be unstable for large N. This illustrates the difficulty of deriving a right fluid model for a system under the proposed scale, which we leave as a future work.

[9] No simulation result is reported in [5] to support this argument.

Acknowledgment

The authors thank the anonymous reviewers for their comments and suggestions that greatly improved the presentation of this paper.

References

[1] S. Floyd, "TCP and explicit congestion notification," *ACM Comput. Commun. Rev.*, vol. 24, no. 5, pp. 10–23, Oct. 1994.

[2] G. Appenzeller, I. Keslassy, and N. McKeown, "Sizing router buffers," in *Proc. ACM SIGCOMM*, Portland, OR, 2004.

[3] J. Sun, M. Zukerman, K. Ko, G. Chen, and S. Chan, "Effect of large buffers on TCP Queueing behavior," in *Proc. IEEE INFOCOM*, Hong Kong, Mar. 2004.

[4] A. Dhamdhere, H. Jiang, and C. Dovrolis, "Buffer sizing for congested Internet links," in *Proc. IEEE INFOCOM*, Miami, FL, Mar. 2005.

[5] G. Raina and D. Wischik, "Buffer sizes for large multiplexers: TCP queueing theory and instability," in *EuroNGI*, Rome, Italy, Apr. 2005.

[6] D. Wischik and N. McKeown, "Buffer sizes for core routers," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 3, pt. I, pp. 75–78, Jul. 2005.

[7] G. Raina, D. Towsley, and D. Wischik, "Control theory for buffer sizing," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 3, pt. II, pp. 79–82, Jul. 2005.

[8] M. Enachescu, Y. Ganjali, A. Goel, N. McKeown, and T. Roughgarden, "Routers with very small buffers," in *Proc. IEEE INFOCOM*, Barcelona, Spain, Apr. 2006.

[9] C. Villamizar and C. Song, "High performance TCP in ANSNET," *ACM Comput. Commun. Rev.*, vol. 24, no. 5, pp. 45–60, 1994.

[10] S. H. Low, F. Paganini, J. Wang, and J. C. Doyle, "Linear stability of TCP/RED and a scalable control," *Comput. Netw.*, vol. 43, no. 5, pp. 633–647, Dec. 2003.

[11] S. Deb and R. Srikant, "Rate-based versus queue-based models of congestion control," in *Proc. ACM Sigmetrics*, New York, NY, Jun. 2004.

[12] D. Hong and D. Lebedev, "Many TCP user asymptotic analysis of the AIMD model," INRIA Tech. Rep. RR-42292001.

[13] P. Tinnakornsrisuphap and A. M. Makowski, "Limit behavior of ECN/RED gateways under a large number of TCP flows," in *Proc. IEEE INFOCOM*, San Francisco, CA, Apr. 2003.

[14] S. Deb and R. Srikant, "Global stability of congestion controllers for the Internet," *IEEE Trans. Autom. Contr.*, vol. 48, no. 6, pp. 1055–1060, Jun. 2003.

[15] S. Shakkottai and R. Srikant, "Mean FDE models for Internet congestion control under a many-flows regime," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1050–1072, Jun. 2004.

[16] F. P. Kelly, "Models for a self-managed Internet," *Philos. Trans. Roy. Soc. A358*, pp. 2335–2348, 2000.

[17] P. E. Lassila and J. T. Virtamo, "Modeling the dynamics of the RED algorithm," in *Proc. QofIS*, Berlin, Germany, 2000.

[18] S. Deb, S. Shakkottai, and R. Srikant, "Stability and convergence of TCP-like congestion controllers in a many-flows regime," in *Proc. IEEE INFOCOM*, San Francisco, CA, Apr. 2003.

[19] P. Tinnakornsrisuphap and R. J. La, "Characterization of queue fluctuations in probabilistic AQM mechanisms," in *Proc. ACM SIGMETRICS*, New York, NY, Jun. 2004.

[20] F. Baccelli, D. R. McDonald, and J. Reynier, "A mean-field model for multiple TCP connections through a buffer implementing RED," *Perform. Eval.*, vol. 49, no. 1-4, pp. 77–97, Sep. 2002.

[21] D. R. McDonald and J. Reynier, "Mean field convergence of a rate model of multiple TCP connections throughput a buffer implementing red," *Ann. Appl. Probab.*, vol. 16, no. 1, pp. 244–294, 2006.

[22] P. Kuusela, P. E. Lassila, and J. T. Virtamo, "Modeling RED with idealized TCP source," in *Proc. IFIP ATM & IP*, 2001.

[23] V. Sharma, J. Virtamo, and P. Lassila, "Performance Analysis of the random early detection algorithm," *Probabil. Eng. Inf. Sci.*, vol. 16, no. 3, pp. 367–388, 2002.

[24] S. M. Ross, *Stochastic Processes*, 2nd ed. New York: Wiley, 1996.

[25] P. Brémaud, *Point Processes and Queues: Martingale Dynamics*. New York: Springer-Verlag, 1981.

[26] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*. New York: Springer-Verlag, 1988.

[27] J. Cao and K. Ramanan, "A Poisson limit for buffer overflow probabilities," in *Proc. IEEE INFOCOM*, New York, NY, 2002.

[28] Y. Zhang, N. Duffield, V. Paxson, and S. Shenker, “On the constancy of Internet path properties,” in *Proc. ACM SIGCOMM Internet Measurement Workshop*, San Diego, CA, Nov. 2001.

[29] H. Jiang and C. Dovrolis, “Why is the Internet traffic bursty in short (sub-RTT) Time Scales,” in *Proc. ACM SIGMETRICS*, Banff, Alberta, Canada, Jun. 2005.

[30] R. Morris and D. Lin, “Variance of aggregated web traffic,” in *Proc. IEEE INFOCOM*, Tel-Aviv, Israel, Apr. 2000.

[31] T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido, “A nonstationary Poisson view of Internet traffic,” in *Proc. IEEE INFOCOM*, Hong Kong, Mar. 2004.

[32] P. Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. New York: Springer-Verlag, 1999.

[33] T. Lindvall, *Lectures on the cOupling Method*. New York: Dover, 2002.

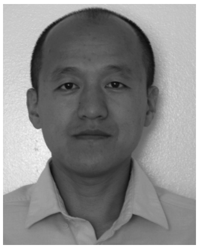
[34] D. Y. Eun and X. Wang, “Achieving 100% throughput in TCP/AQM under aggressive packet marking with small buffer,” North Carolina State Univ., Raleigh, NC, Tech. Rep., Aug. 2006 [Online]. Available: http://www4.ncsu.edu~dyeun/pub/techrep06-smallbuffer.pdf

[35] *ns*-2 — The Network Simulator. ISI, 2004 [Online]. Available: http://www.isi.edunsnamns

Do Young Eun (M’03) received the B.S. and M.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Taejon, Korea, in 1995 and 1997, respectively, and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, in 2003.

Since August 2003, he has been an Assistant Professor with the Department of Electrical and Computer Engineering at North Carolina State University, Raleigh. His research interests include modeling and analysis of wireless networks, congestion control, resource allocation, mobility modeling, and ad hoc/sensor networks.

Dr. Eun is a member of Technical Program Committee of IEEE INFOCOM 2005–2008, IEEE ICC 2005–2007, IEEE Globecom 2005, and IEEE IPCCC 2006, 2007, among others. He received the Best Paper Awards in the IEEE ICCCN 2005 and the IEEE IPCCC 2006, and the NSF CAREER Award 2006. He is a member of ACM.



Xinbing Wang (M’06) received the B.S. degree (with hons.) from the Department of Automation, Shanghai Jiaotong University, Shanghai, China, in 1998, and the M.S. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2001. He received the Ph.D. degree, major in the Department of Electrical and Computer Engineering, minor in the Department of Mathematics, North Carolina State University, Raleigh, in 2006.

Currently, he is a faculty member in the Department of Electronic Engineering, Shanghai Jiaotong University, Shanghai, China. His research interests include resource allocation and management in mobile and wireless networks, TCP asymptotics analysis, wireless capacity, cross layer call admission control, asymptotics analysis of hybrid systems, and congestion control over wireless ad hoc and sensor networks.

Dr. Wang has been a member of the Technical Program Committees of several conferences including IEEE ICC 2007, IEEE Globecom 2007, IEEE WCNC 2007, IEEE ICCCN 2007 and IEEE IPCCC 2007. He is a member of the ACM.