Bandwidth Management for Mobile Media Delivery

Sanjeev Mehrotra [#], Hua Chen [*], Sourabh Jain [†], Jin Li [#], Baochun Li [◇], Minghua Chen [○]
[#] *Microsoft Research* – {sanjeevm,jinl}@microsoft.com [*] *University of Maryland* – huachen@umd.edu
[†] *University of Minnesota* – sourj@cs.umn.edu [◇] *University of Toronto* – bli@eecg.toronto.edu
[○] *The Chinese University of Hong Kong* – minghua@ie.cuhk.edu.hk

***Abstract*—Mobile broadband networks using 3G and 4G technologies (such as EV-DO, HSPA, WiMAX, LTE) are rapidly becoming one of the prominent means to access the Internet. Multimedia consumption — requiring low delay, high bandwidth, or a combination of both — is projected to become a large portion of bandwidth utilization in mobile broadband networks. In this paper, we study the fundamental problem of how packet loss and delay vary as a function of the transmission rate over these networks. With extensive real-world measurement studies, we analyze the performance of a number of rate control algorithms commonly used in media transmission. We show that the variable nature of congestion signals (loss and delay) on these networks leads to an ultimate failure of existing rate control strategies to deliver adequate performance for multimedia applications. In addition, we show how a rate control algorithm derived from the utility maximization framework — which uses queuing delay as the primary congestion signal — can be modified to solve the challenging issues we have observed. By using a variable threshold to define when the network is congested, our proposed solution is able to achieve a significant improvement over algorithms that use fixed definitions of congestion.**

I. Introduction

Mobile broadband Internet usage is rapidly on the rise. In the US and Europe, the proportion of mobile users that access the Internet regularly has quadrupled and tripled over the past four years, respectively [5]. Owners of smartphones are driving the increase in mobile Internet usage, and most consumers access the Web on their smartphones primarily via their cellular connection, rather than Wi-Fi. As a result, 3G cellular technology is driving the growth of mobile broadband usage and is currently available to more than 20 percent of cellular users around the world. 4G mobile broadband technologies, such as WiMAX and LTE, are being quickly deployed as well. Among smartphone activities, more and more users are utilizing bandwidth intensive applications. Mobile video is experiencing explosive growth and is driving the growth of mobile broadband. According to [17], mobile video has represented 47% of peak hour traffic in November 2010, up from 27% in January 2010. In addition, interactive media applications — such as video conferencing on mobile broadband — are quickly becoming popular. Applications such as *FaceTime* and *Skype* are widely available on smartphones, and are redefining the mobile video conferencing experience.

Depending on the interaction between mobile devices and the cloud, there are primarily two categories of mobile broadband applications: *streaming* and *interactive*. For streaming applications, such as map browsing and video-on-demand (e.g., YouTube, NetFlix, Hulu), the effective throughput determines the quality of experience. Other network parameters, such as network latency (propagation delay), queuing delay, and packet loss, are less relevant. These applications either have a typical response time of several seconds (in map browsing), or build a buffer of several seconds (in video-on-demand apps) at the mobile devices to absorb the queuing delay and packet loss. On the other hand, interactive applications — such as games and video conferencing — strive to achieve a good throughput with queuing delay and packet loss.

In our discussion, by *queuing delay*, we mean the the delay experienced by a packet waiting somewhere along the network path and is defined as the delay minus the minimum delay. It is also often referred to as either *jitter* or *packet delay variation*. Some portion of *packet loss* and *queuing delay* can be attributed to congestion and we refer to that as *congestion induced*.

A. Related Work

There exists an extensive body of literature on rate and congestion control in the Internet. However, few addresses the unique issue of bandwidth management for media delivery on mobile broadband networks. There are two major categories of congestion control and rate control algorithms in the literature: (1) those based on the estimation of available bandwidth, and (2) those based on end-to-end congestion control.

An example of rate control based on available bandwidth estimation is [18]. However, it has been shown [11] that this category of technologies frequently fail in complex networks and will have issues in an environment such as mobile broadband where both loss and delay measurements have significant noise. Moreover, such approaches fail to share bandwidth fairly and thus we do not consider them.

End-to-end congestion control, such as the standard Additive-Increase Multiplicative-Decrease (AIMD) algorithm used in TCP [9], dominates the Internet. There have been many variants of TCP developed over the years – [4], [8], [9], [14], [19] are just some of the examples.

Loss based end-to-end congestion avoidance algorithms such as [8], [9] work relatively well over *wired Internet connections* where loss is primarily due to congestion and for *non-interactive* applications such as video streaming or file download where only *throughput* is important. In particular they do not work well when there is random packet loss on the network as is the case for mobile broadband networks or for interactive applications where *queuing delay* is also important for application performance in addition to throughput.

Delay based congestion avoidance algorithms such as [4] and combined loss/delay based algorithms such as [19] attempt to minimize queuing delay and thus may work better for interactive applications. However, as they only use *fixed definitions of congestion*, there will always be certain networks for which these algorithms suffer from link underutilization or operate at a congestion level which is too high for interactive applications to work well.

There have also been several attempts to improve the performance of congestion control protocols over networks with wireless links, where not all loss is caused by congestion. A good summary of existing work is provided in [2]. However, existing work has mostly focused on *loss-based* protocols, with the goal being to either hide loss from the upper layer (such as TCP) via retransmissions at the lower layer [1], [3], or modify the upper layer protocol to determine whether observed loss is actually due to congestion [4], [14], [19]. The former approach requires changes in the hardware or firmware which is difficult to implement. As an example of the latter approach, TCP Westwood [14] attempts to determine which loss is due to congestion by utilizing bandwidth estimation techniques to set the slow start threshold and initial congestion window. Although this works well when only *loss* is a noisy congestion signal, it does not work well on mobile broadband networks where *queuing delay* is also a noisy signal since bandwidth estimation techniques themselves do not work well.

B. Contributions

To address the challenge of high bandwidth and low queuing delay and packet loss needed by multimedia applications, our original contribution in this paper is improving a queuing delay based rate control algorithm inspired from the utility maximization framework [12], [13]. However, as opposed to most existing rate control strategies which use a *fixed definition of congestion*, we propose to use an *adaptive definition of congestion* which allows us to achieve close to full throughput but without incurring additional queuing delay or packet loss. This is essential in improving the performance of multimedia applications over mobile broadband networks.

In this paper, we revisit the fundamental challenge of determining when observed congestion signals — such as queuing delay and packet loss — are in fact due to congestion, but over mobile broadband networks where these signals are inherently *noisy*. We make the following contributions.

- In Sec.II, we show that mobile broadband networks have very high inherent queuing delay and packet loss levels as opposed to traditional networks. This occurs even in the absence of congestion.
- In Sec. III, we show that existing rate control algorithms using fixed *definitions of congestion* perform relatively poorly in terms of throughput and/or packet loss and queuing delay. By *definition of congestion*, we mean the queuing delay or packet loss boundary (the threshold) between the *congestion* and *congestion-free* zone.
- In Sec. IV, we propose a rate control algorithm using a *variable* definition of congestion as opposed to a *fixed* definition, which has not been proposed in the literature. Although using a high fixed threshold may result in full link utilization for most networks, it will have a detrimental effect on real-time (interactive) applications where delay is also important. We show the performance of the algorithm in Sec. V.

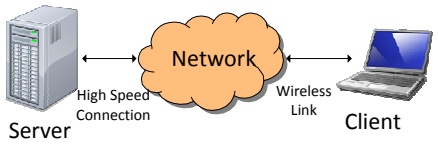


Fig. 1. The system setup for network trace collection. The server is connected to the Internet via a high-speed connection. The client is connected to the Internet via a mobile broadband connection such as 3G or WiMAX.

Our rate control algorithm differs along several directions with respect to previous work.

- We deal with congestion control protocols which also use *queuing delay* as primary congestion signals as opposed to just *packet loss*.
- We deal with networks which have *variable levels* of noise in queuing delay and loss as opposed to *fixed levels*.
- Rather than detecting if observed delay or loss is actually due to congestion, we attempt to learn inherent queuing delay and loss levels in the congestion-free zone and design a rate control strategy which works around them.

Although we present results for a couple of representative mobile broadband networks — specifically for a 3G EV-DO network (referred to as simply 3G) and for a WiMAX network – we find that the contributions of this paper hold in most other mobile broadband networks as well.

II. Empirical Observations from Network Traces

We first present results from extensive network traces collected from 3G and WiMAX networks. Our network trace collection is performed using the setup in Fig. 1, where the server is connected to the Internet via a high-speed Internet link, and the client is connected via a mobile broadband link such as 3G or WiMAX. In this setup, the most likely bottleneck link is the mobile broadband link. We refer to "upload" traffic when the client sends data to the server and "download" traffic when the server is sending to the client. We send packets in the upload or download direction using a payload size of M bytes at a rate of R bytes/sec. The receiving end observes the arriving packets and records queuing delay and loss measurements. The rate R is varied over various transmission rates, and packets are sent at each rate for a duration of 20 seconds before moving to the next rate. We have collected traces in both download and upload directions and at different times of the day as well as for various packet sizes. The total traces span the course of data collected for one hour per day over the course of one week.

In the figures, we show the queuing delay and loss rate observed when sending packets at various rates. The x-axis shows the rate in kbps and the y-axis shows the one-way queuing delay (OWD) in seconds for the delay plots, and the loss rate as a fraction for the loss plots. The OWD is estimated

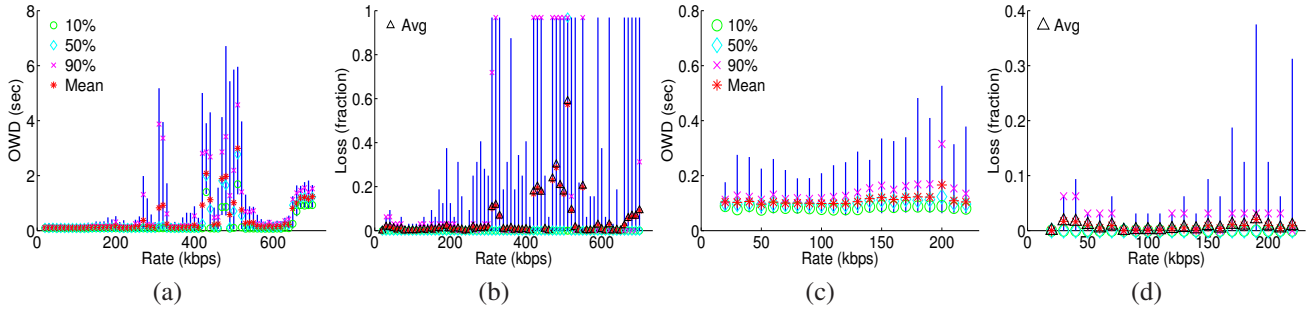


Fig. 2. Results for 3G network in upload direction showing (a) one-way queuing delay (OWD) and (b) loss rate for the 3G network in upload direction as a function of transmission rate. In (c), we show enlarged version of the OWD results in the congestion-free zone, and in (d), we show the loss rate in the congestion-free zone. We show the range of values observed, with the 10%, 50%, 90% and mean marked. The loss rate shown is computed over a sliding window of 32 packets. The overall average loss rate at a particular bitrate is also marked.

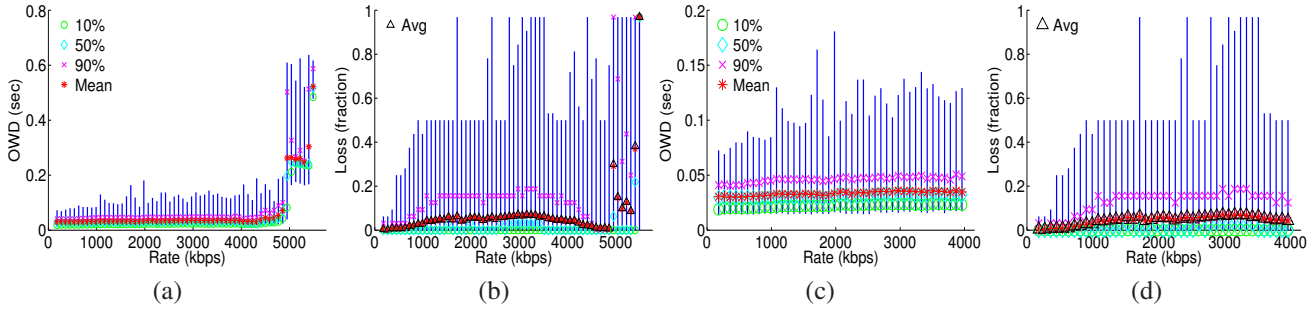


Fig. 3. Same results as in Fig. 2, but for WiMAX network in the download direction.

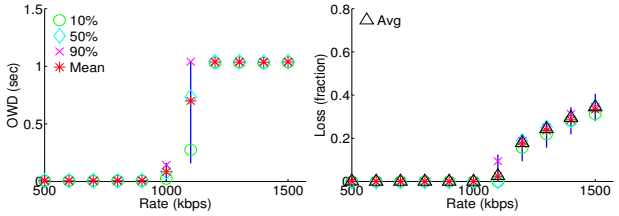


Fig. 4. Statistics for a cable modem connection in upload direction.

as one-half the RTT, although using clock drift compensation techniques, true OWD values can also be obtained. At each particular transmission rate, we draw a straight vertical line, with the bottom of the line representing the minimum queuing delay observed and the top of the line representing the maximum queuing delay observed at that rate. We use a circle, a diamond, and a cross to represent the 10-, 50- (median), and 90-percentile queuing delay observed, respectively, and use a star to represent the average queuing delay at each rate. We visualize the loss data in a similar fashion, except each loss data point is calculated from the packet loss rate in a sliding window of 32 packets. Thus, if we have sent $1,000$ packets at a particular rate, we will have $1000 - 32 = 968$ data points.

We show network characteristics for the 3G network in upload direction (Fig. 2) and for the WiMAX network in download direction (Fig. 3). These results use a packet size of 1000 bytes. In the figures, we also show enlarged versions of the queuing delay and loss rate for the rates in the "congestion-free" zone. For comparison, we also show network characteristics of a cable modem link in Fig. 4. A larger set of results with the direction of traffic being reversed, varying packet sizes, and varying times of day can be found in [15].

We first observe that for each network (3G, WiMAX, and cable modem), there exists a clear distinction between the *congestion-free* zone and the *congestion* zone. That is, there exists a rate above which the loss and queuing delay statistics clearly start to rise. Thus, there exists a metric, which could be loss, queuing delay, or some combination, that can be used to detect congestion.

Our second observation is that the 3G and WiMAX networks exhibit much more variation in terms of queuing delay and loss in the congestion-free zone than the cable modem link. For the cable modem link, any queuing delay above a small amount (such as 50 ms) can be considered congestion and any loss can be considered congestion as well. For 3G and WiMAX networks, significant variations of loss and queuing delay may be observed even when the network is in the congestion-free zone. For example, we see 2-3% packet loss even in the congestion-free case, along with 50-100 ms queuing delays.

We have repeated these measurements using reverse traffic directions, varying packet sizes, and at different times of the day (for example at night). Although the overall observed bandwidth may be slightly different — for example greater bandwidth in the download direction or at night — the inherent noise in queuing delay and packet loss remains.

III. Existing Rate Control Protocols

We compare the performance (in terms of throughput, delay, and loss) of several media rate control algorithms designed from the principles of existing congestion control protocols. We show that *varying* levels of inherent queuing delay and

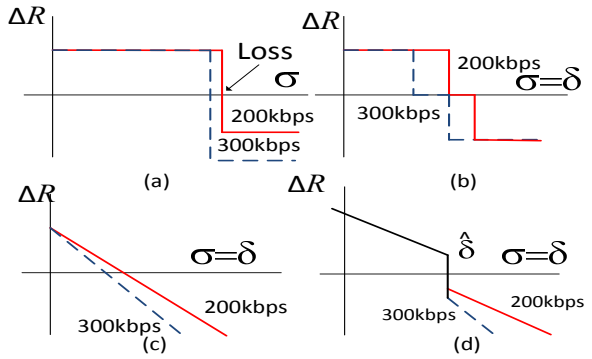


Fig. 5. Change in rate (ΔR) vs. congestion level (σ) for various congestion control algorithms for two different rates, (a) TCP NewReno-like rate control, (b) TCP Vegas-like rate control, (c) Primal-Dual Utility Maximization (UM), (d) Our solution presented in Sec. IV. Both (c) and (d) can easily be modified to have variable congestion thresholds using our strategies from Sec. IV-B.

TABLE I
RESULTS ON 3G NETWORK UPLOAD – "ADAPTIVE" REFERS TO THE ALGORITHM IN SEC. IV.

	Throughput (kbps)	Delay (sec)	Loss (fraction)
NewReno-like	164	0.12	0.041
TFRC	198	0.10	0.104
Vegas-like	209	0.13	0.007
CTCP-like	223	0.15	0.478
UM	219	0.45	0.744
Adaptive	207	0.10	0.062

TABLE II
RESULTS ON WIMAX NETWORK DOWNLOAD.

	Throughput (kbps)	Delay (sec)	Loss (fraction)
NewReno-like	585	0.018	0.019
TFRC	1154	0.017	0.044
Vegas-like	578	0.017	0.012
CTCP-like	2629	0.085	0.097
UM	2285	0.020	0.470
Adaptive	3901	0.042	0.059

loss in mobile broadband networks makes it very difficult for existing rate control algorithms which used *fixed definitions of congestion* to obtain adequate performance regardless of whether delay, loss, or both delay and loss are used as congestion signals. We evaluate several TCP like protocols [4], [7], [9], [19] and delay-based primal-dual utility maximization (UM) with fixed parameters [6]. We don't consider available bandwidth estimation techniques [10], as they provide no guarantees of fairness across multiple flows and are known to perform poorly on noisy networks. In the following discussion, we use the following notation: R: the transmission rate, W: TCP congestion window, δ: observed queuing delay (observed delay minus minimum delay), and ϵ: observed loss rate.

Using the traces from Sec. II, for the 3G and WiMAX network, we simulate the performance of these rate control protocols. For each protocol, we compute the overall transmission rate (throughput), queuing delay, and loss rate. The loss rate is computed using a sliding window as explained before. The results are summarized in Tables I and II. The throughput reported is the rate which is achieved accounting for packet loss ("goodput"). Detailed figures showing rate as a function of time and PDF of delay and loss rate can be found in [15]. For brevity, results are shown for the 3G network in the upload direction and for the WiMAX network in the download direction. Other configurations show similar results.

A. TCP Variants

In TCP NewReno-like rate control, we use a method inspired from TCP AIMD congestion control [9]. TCP increments the window by $\frac{M^2}{W}$ for every ACK received and decrements the window by $-\frac{W}{2}$ for every NACK, where M is the packet size. Since media applications typically use rate control instead of window control, we translate the window to transmission rate using $W = R \cdot SRTT$, where $SRTT$ is the smoothed round-trip time (RTT). We increment the rate by $\frac{M^2}{RSRTT^2}$ for every ACK, and decrement the rate by $-\frac{R}{2}$ for every NACK. The rate change curve vs. congestion level is shown in Fig. 5(a).

When using TCP NewReno-like rate control, we expect that for most networks it results in full network utilization, but at large queuing delays and some amount of congestion induced packet loss. However, for mobile broadband networks, we actually see significant link underutilization (Tables I, II). We see that the 3G link is only about 50% utilized and the WiMAX link is only about 12% utilized. This link under-utilization is caused by the fact that in these networks, there is significant random loss, and thus a single packet loss by itself is not necessarily congestion as loss based TCP assumes.

From Tables I and II, we see TFRC exhibits similar throughput, delay, and loss as TCP NewReno-like rate control. This is understandable as TFRC uses the TCP NewReno throughput equation to set the transmission rate [7].

Since loss is not a good congestion signal, we consider a delay based rate control algorithm, similar in spirit to TCP Vegas [4], which uses queuing delay above some threshold as a signal of congestion. In TCP Vegas-like rate control, for each ACK we increase the rate by $\frac{M^2}{R \cdot SRTT^2}$ if $\delta < \kappa$, do nothing if $\kappa \leq \delta < \zeta$, and decrease the rate by $\frac{M^2}{R \cdot SRTT^2}$ if $\delta \geq \zeta$, where κ and ζ are constants. Upon NACK, we decrease the rate by $-\frac{R}{2}$. The rate change curve is shown in Fig. 5(b).

From Table I, we see better link utilization than the TCP NewReno-like rate control over a 3G network. However, from Table II, we still see link under-utilization for the WiMAX network in the download direction. Although delay may be a good signal of congestion, the use of fixed parameters makes it suitable for only one type of network. For example, if we optimize the parameters to work well in a 3G network in the upload direction at a particular bitrate, it may sill not work effectively in the WiMAX network in the download direction (still results in link under-utilization) because of the operating point not being correct. Rate control algorithms with fixed parameters have an operating congestion point which decreases with bitrate. Since the WiMAX network in the download direction has much higher bitrate than the 3G network in the upload direction, the delay operating point is much lower for the WiMAX network in the download direction. Since it is within the inherent noise in delay for this network, it causes link under-utilization.

We also evaluate a rate control inspired by a recent TCP variant, Compound TCP [19], which was developed for use

on high bandwidth-delay product networks which uses both delay and loss. Although we see that the link is fairly well utilized in both cases, the use of fixed parameters results in too high of an operating congestion point on the 3G network (over 40% loss), which is unacceptable.

B. Delay Based Primal-Dual Utility Maximization

In primal-dual utility maximization (UM) using the log utility function, we attempt to maximize the total utility over the network, in which each source has a utility, U, vs. rate, R, given by $U(R) = k_0 \log(R)$, where k_0 is a constant. Provided the buffers along the path are of a sufficient size (larger than the operating congestion point, $\hat{\delta}$, defined below), this gives a rate control algorithm that adjusts the rate using

$$\Delta R = k_2(k_0 - \delta R)\Delta T, \tag{1}$$

where δ is the observed queuing delay, k_2 is a constant, and ΔT is the time since the previous adjustment. k_2 controls the rate of convergence and steady-state oscillation. This rate change curve is shown in Fig. 5(c).

The operating congestion point is given by the δ where $\Delta R = 0$. From (1), the operating congestion point is a queuing delay of $\hat{\delta} = \frac{k_0}{R}$, where R is the steady state rate. If we fix k_0, then we see that $\hat{\delta}$ is a function of rate. For networks where delay is a noisy signal, if $\hat{\delta}$ is set too low (for example if k_0 is chosen too low), then the observed queuing delay may be erroneously classified as due to congestion and the link will be under-utilized. This would result in poor performance for bandwidth intensive multimedia applications. such as video conferencing or video on demand. Choosing a high k_0 and thus a high $\hat{\delta}$ may provide full link utilization, but will result in poor performance (in terms of queuing delay and packet loss) for real-time and interactive multimedia applications.

From Tables I and II, we see that improperly selecting the parameter k_0 in the UM algorithm has resulted in too high of an operating congestion point (too much packet loss), even though the link is fairly well utilized. We note that regardless of the k_0 used, there may be some network where the choice is either too high or too low.

IV. Alternative Rate Control Design

A. Congestion Signals

In Sec. II, we have seen that queuing delay and loss are noisy congestion signals on mobile broadband networks, and in Sec. III, we have seen that using fixed definitions of congestion result in poor performance. We naturally ask the question: for 3G and WiMAX networks, what is the best way to distinguish whether we are in the congestion zone or in the congestion-free zone? Should the congestion signal be *queuing delay*, *packet loss*, or some combination of the two? In order to make an informed decision, we attempt to classify the queuing delay and loss measurements in Figs. 2 and 3 into a congestion zone and a congestion-free zone.

For each network, we define a set of "uncongested rates" as those rates where the historical average delay and loss seen is no larger than some percentage of the average delay and loss seen when transmitting at a very low rate (for example no larger than 1.2x of the delay/loss seen when transmitting at 20kbps). We note that of course at any particular moment, even transmitting at these rates may actually be congesting the link. However, this analysis is done from a large set of measurements and is only meant for determining appropriate congestion signals for a particular network. Similarly, we define a set of "congested rates" as those where the historical average delay or loss seen is larger than some percentage of the average delay and loss seen when transmitting at a very low rate (for example larger than 2.0x of the delay/loss seen when transmitting at 20kbps). We define R_{max}^{uncong} to be the maximal rate in the set of uncongested rates and R_{min}^{cong} to the minimal rate in the set of congested rates. As an example, for the 3G network in the upload direction in Fig. 2, we find $R_{max}^{uncong} = 250$kbps and $R_{min}^{cong} = 640$kbps. For the WiMAX network in the download direction in Fig. 3, we find $R_{max}^{uncong} = 4900$kbps and $R_{min}^{cong} = 5100$kbps.

Using the definitions of R_{max}^{uncong} and R_{min}^{cong} above, we compute the PDF of delay for the uncongested and congested measurements as $P(\delta | R \leq R_{max}^{uncong})$ and $P(\delta | R \geq R_{min}^{cong})$, where $P(.)$ represents the PDF. We similarly compute the PDF of loss in these two sets and show the results in Fig. 6.

We observe that in both 3G and WiMAX broadband networks, delay is a good signal for congestion detection. We can pick a delay threshold, δ_T, such that most values from the "uncongested rates" measurement set fall below δ_T and most values from the "congested rates" measurement set fall above. That is we can find a δ_T which can be used to classify the region of congestion vs. non-congestion so that $P(R \leq R_{max}^{uncong} | \delta \leq \delta_T)$ and $P(R \geq R_{min}^{cong} | \delta > \delta_T)$ are high. On the other hand, loss is not a good signal for congestion detection. For example on the 3G network, loss characteristics are very similar for both the congestion-free and congested set. In either zone, we may see 2% packet loss. For the WiMAX network, we can choose a loss threshold, ϵ_T such that the two sets are separated. However, even a 5% loss rate is fairly common in the congestion-free zone and thus the threshold needed is very large (20% loss rate) which is prohibitively high for real-time applications. Clearly, if we use any loss to be taken as congestion, as TCP NewReno-like rate control does, we under-utilize the link.

B. Rate Control Using Variable Definition of Congestion

Since delay is an appropriate indicator of congestion, we start from the delay based UM rate control framework in Sec. III-B. We show that using variable definitions of congestion can result in good link utilization as well as low queuing delay and packet loss. In (1), the operating congestion point, $\hat{\delta}$ is controlled through k_0. A modified version is presented in [16], which directly controls the operating congestion point $\hat{\delta}$,

$$\Delta R = \begin{cases} \alpha_\delta & \text{if } \delta \leq \hat{\delta} \\ -\beta_\delta R & \text{if } \delta > \hat{\delta} \end{cases}, \tag{2}$$

where α_δ and β_δ are fixed constants or functions of δ used in the AIMD adjustment. For example, α_δ and β_δ can be linear

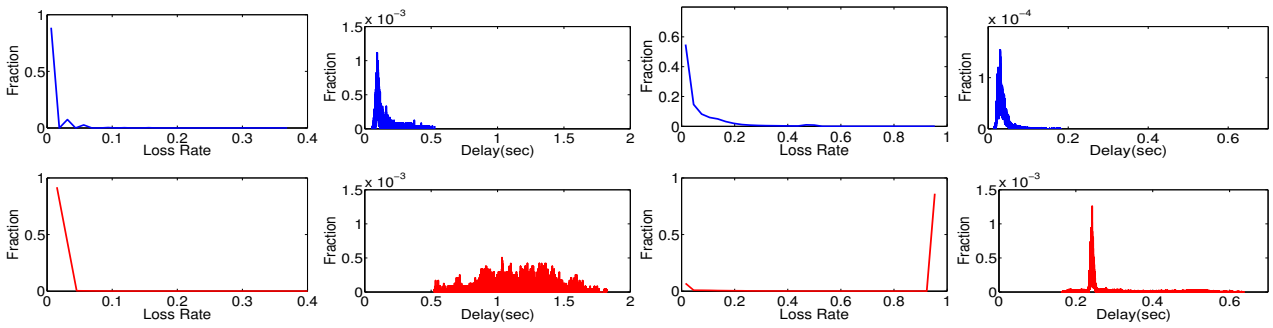


Fig. 6. The PDF in the congestion-free zone (top figure) and the congestion zone (bottom figure) for (a) loss rate over 3G network upload, (b) delay over 3G network upload, (c) loss rate over WiMAX network download, and (d) delay over WiMAX network download.

functions of δ as shown in Fig. 5(d): $\alpha_\delta = \alpha_{\max} + (\alpha_{\min} - \alpha_{\max})\frac{\delta}{\hat{\delta}}$, $\beta_\delta = \min\left(\beta_{\max}, \beta_{\min} + (\beta_{\max} - \beta_{\min})\frac{\delta - \hat{\delta}}{\delta_{\max} - \delta}\right)$. Regardless of whether (1) or (2) is used, the operating congestion point is given by $\hat{\delta}$ and is either directly set as in (2) or indirectly set through k_0 as in (1). In order for the rate control strategy to provide the lowest possible congestion point given network characteristics while providing full link utilization, we need to find the *lowest* $\hat{\delta}$ to properly disambiguate the boundary between the *congestion* zone and the *congestion-free* zone.

For full link utilization, it is essential to not confuse inherent noise in congestion signals as actual congestion. In order to accomplish this, we can choose the operating point of the congestion algorithm, $\hat{\delta}$ as $\hat{\delta} \geq \delta_T$, where δ_T is the boundary between the congestion zone and the congestion-free zone as described in Sec. IV-A. We use the following to set the operating point in (2):

$$\hat{\delta} = \theta\delta_T \geq \delta_T = E[\delta | R \in R^{uncong}], \tag{3}$$

where $\theta > 1 = 1.25$ and $R^{uncong} = \{R | \delta_{avg}(R) \leq \alpha\delta_{avg}(R^{base})\}$, where $\alpha = 1.1$ and $R^{base} = 20$kbps. $E[\delta | R \in R^{uncong}]$ is the average delay seen in the "uncongested set of rates" as defined in Sec. IV-A. The operating point is set to something slightly larger than the average uncongested delay.

V. Performance

We show the performance of the UM based rate control with variable definitions of congestion. We use the rate control algorithm in Eqn. 2 with $\hat{\delta} = 0.4$ for the 3G network and $\hat{\delta} = 0.1$ for the WiMAX network which is found using (3). We compare the throughput, delay, and loss rate with existing schemes presented in Sec. III in Tables I and II. We also show these results in Figs. 7 and 8.

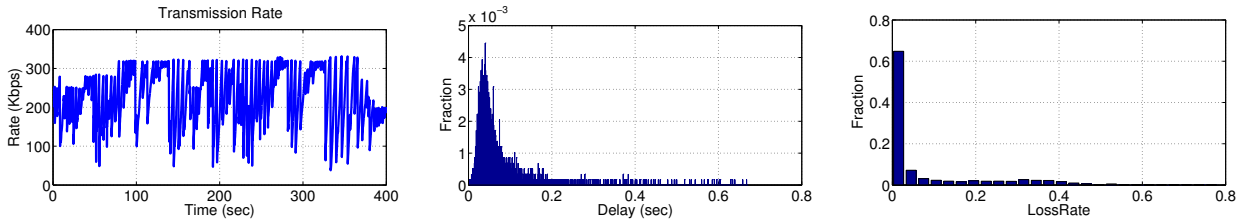
Compared with TCP-NewReno-like rate control, TCP-Vegas-like rate control, CTCP-like rate control and TFRC, we see that by learning the proper congestion operating point, we are able to achieve good link utilization at reasonable delay and loss levels, which are close to the 70-80th percentile of the inherent noise level within the network. We see that for the 3G network, we have close to 26% of throughput gain compared to TCP-NewReno-like rate control and similar throughput as CTCP-like rate control but with much lower loss. For the WiMAX network, we see 50% throughput improvement over CTCP-like rate control and a UM algorithm with fixed parameters with lower loss and delay; and a 7x improvement in throughput over TCP-NewReno-like and Vegas-like rate control.

In Fig. 9, we show fairness of the protocol across two flows. One flow is run from 0-1000 seconds, and a second parallel flow is run from 200-700 seconds. During the period where one flow is running, it obtains approximately the full 5Mbps bandwidth, with a delay of about 20ms. When the second flow enters, they both take about 2.5Mbps each, with the queuing delay remaining close to 20ms. Additionally, unlike standard rate control algorithms with fixed parameters, the delay does not increase when the transmission rate drops. Instead it remains close to the minimum allowed by the inherent noise in the network. This is because our rate control framework tries to achieve a desired operating delay point rather than using fixed parameters. The loss for the entire session is also negligible and stays close to 0.2%. Although convergence currently takes about 20-30 seconds, techniques similar to that used in TCP slow start can be used to speed up convergence.

In Fig. 10, we show the throughput (R) and delay (δ) achieved when using different values of $\hat{\delta}$ in (2). We see that if we choose $\hat{\delta}$ appropriately and larger than δ_T, we can achieve full throughput. However, in order to keep the operating congestion point (queuing delay) low, we should choose $\hat{\delta}$ as small as is needed for full throughput by using (3). Since δ_T is a function of the network, we see that $\hat{\delta}$ also needs to adapt to the network.

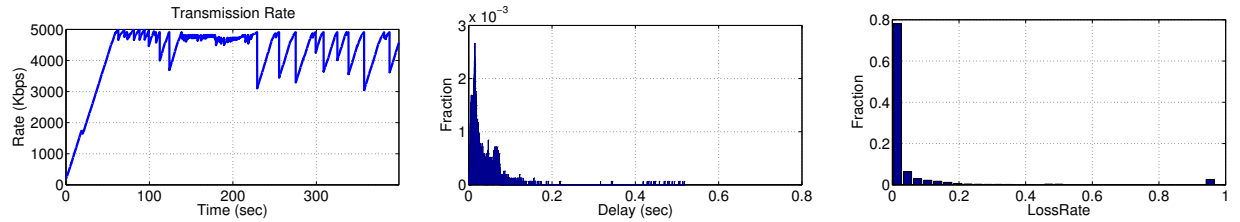
VI. Conclusion

In this paper, through the lens of extensive illustrative network traces, we have shown strong evidence that loss and delay are often very noisy signals in mobile broadband networks. When this is the case, existing rate control techniques — such as TCP like rate control schemes, TFRC, and other utility maximization schemes with fixed parameters — either fail to fully utilize the link or fail to function at reasonable queuing delay and loss operating points. This makes media delivery and especially real-time video conferencing very problematic in such networks. The highlight of this paper is our demonstration



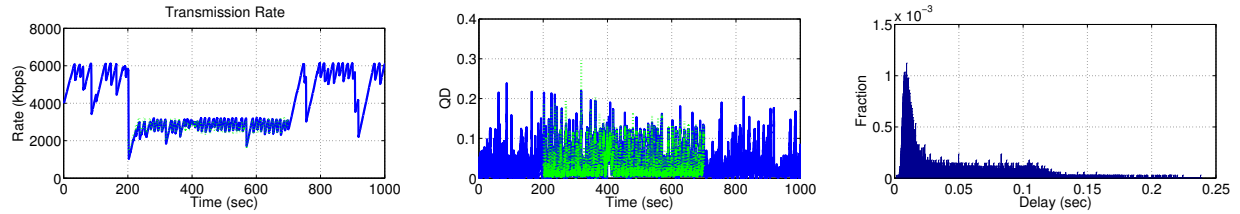
(a) (b) (c)

Fig. 7. (a) Transmission rate, (b) PDF of queuing delay, and (c) PDF of loss rate when using learning based algorithm on 3G upload.



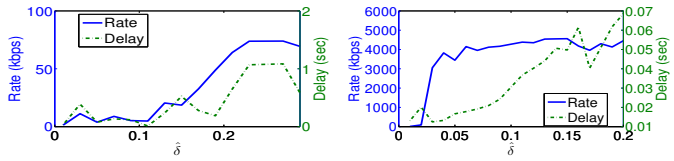
(a) (b) (c)

Fig. 8. (a) Transmission rate, (b) PDF of queuing delay, and (c) PDF of loss rate when using learning based algorithm on WiMAX download.



(a) (b) (c)

Fig. 9. (a) Transmission rate across time, (b) queuing delay across time, and (c) PDF of queuing delay when using a learning based algorithm on WiMAX network with two flows. Two parallel flows are running from 200-700 sec. Overall loss rate is about 0.2%.



(a) (b)

Fig. 10. Rate and operating congestion level (observed queuing delay) as function of $\hat{\delta}$ for (a) 3G upload and (b) WiMAX download

that the state of congestion can be *learned* relatively easily in mobile broadband networks, and our proposal of a new adaptive algorithm to effectively learn congestion operating points, with significantly improved performance in our real-world experiments.

REFERENCES

[1] E. Ayanoglu, S. Paul, T. F. LaPorta, K. K. Sabnani, and R. D. Gitlin. AIRMAIL: A link-layer protocol for wireless network. *ACM Wireless Networks*, Feb. 1995.

[2] H. Balakrishnan, V. Padmanabhan, S. Seshan, and R. Katz. A comparison of mechanisms for improving TCP performance over wireless links. *IEEE/ACM Trans. Networking*, 5(6):756–769, 1997.

[3] H. Balakrishnan, S. Seshan, and R. Kat. Improving reliable transport and handoff performance in cellular wireless networks. *ACM Wireless Networks*, Dec. 1995.

[4] L. Brakmo, S. O'Malley, and L. Peterson. TCP Vegas: New techniques for congestion detection and avoidance. In *Proc. ACM SIGCOMM*, pages 24–35, Aug. 1994.

[5] P. Brown. Mobile internet usage continues to rise: Android users catching up with Apple. *Strategy Analytics*. Jan. 2011.

[6] M. Chen, M. Ponec, S. Sengupta, J. Li, and P. A. Chou. Utility Maximization in Peer-to-Peer Systems. *Microsoft Research Technical Report*, August 2007.

[7] S. Floyd, M. Handley, J. Padhye, and J. Widmer. Equation-based congestion control for unicast applications. In *Proc. ACM SIGCOMM*, pages 43–56, Stockholm, Sweden, Aug. 2000.

[8] S. Ha, I. Rhee, and L. Xu. CUBIC: A new TCP-friendly high-speed TCP variant. *ACM SIGOPS Operating System Review*, 42(5):64–74, Jul 2008.

[9] V. Jacobson. Congestion avoidance and control. In *Proc. ACM SIGCOMM*, pages 314–329, Stanford, CA, Aug. 1988.

[10] M. Jain and C. Dovrolis. End-to-end available bandwidth: Measurement methodology, dynamics, and relation with TCP throughput. *IEEE/ACM Trans. Networking*, 11:537–549, Aug. 2003.

[11] M. Jain and C. Dovrolis. Ten fallacies and pitfalls on end-to-end available bandwidth estimation. In *IMC*, 2004.

[12] F. P. Kelly, A. Maulloo, and D. Tan. Rate control for communication networks: shadow prices, proportional fairness, and stability. *Journal of the Operationl Research Society*, 49:237–252, 1998.

[13] S. H. Low and D. E. Lapsley. Optimization flow control, i: Basic algorithm and convergence. *IEEE/ACM Trans. Networking*, 7(6):861–875, Dec. 1999.

[14] S. Mascolo, C. Casetti, M. Gerla, M. Y. Sanadidi, and R. Wang. TCP Westwood: Bandwidth estimation for enhanced transport over wireless links. In *Proceedings of the 7th annual international conference on Mobile computing and networking*, MobiCom '01, pages 287–297, New York, NY, USA, 2001. ACM.

[15] S. Mehrotra, H. Chen, S. Jain, J. Li, B. Li, and M. Chen. Bandwidth management for mobile media delivery. Technical Report MSR-TR-2011-105, Microsoft Research, July 2011.

[16] S. Mehrotra, J. Li, S. Sengupta, M. Jain, and S. Sen. Hybrid window and rate based congestion control for delay sensitive applications. In *Proc. of IEEE Globecom*. IEEE, Dec. 2010.

[17] M. Murphy and M. Meeker. Top mobile internet trends. KPCB Relationship Capital. Feb. 2011.

[18] J. Strauss, D. Katabi, and F. Kaashoek. A measurement study of available bandwidth estimation tools. In *IMC*, Oct 2003.

[19] K. Tan, J. Song, Q. Zhang, and M. Sridharan. A compound TCP approach for high-speed and long distance networks. In *INFOCOM*, pages 1 – 12. IEEE, Apr. 2006.