

The impact of link layer assisted multimedia adaptation in wireless networks

Dr. Manthos Kazantzidis, Dr. Mario Gerla
{kazantz, gerla}@cs.ucla.edu

Abstract-- Higher layer protocols in wireless networks need to dynamically adapt to observed network response. A common approach in multimedia is to employ end-to-end monitoring to estimate quantities of interest like delay, delay jitter and available bandwidth. In this paper we show by experimentation that end-to-end adaptation does not work in networks with wireless 802.11 links. Transport measurements are unreliable and converge very slowly. Therefore we develop a link layer assisted architecture for one and multi hop wireless networks. We first show that end-to-end mechanisms cannot scale to many connections and it is preferable to merely use low rate non adaptive streams. On the other hand our link-network feedback architecture is very successful and scalable to number of connections. It offers increased QoS when this is feasible, or otherwise at least the QoS of a low rate transmission.

Index terms-- Network architectures, Network assisted congestion control, link assisted congestion control, congestion control, Quality of service, Ad hoc & sensor networks, System design, Simulations, Network measurements

A. INTRODUCTION

A great deal of work is targeted at exploiting adaptive mechanisms in all design layers of wireless networks. The goal is to gain the desired protocol responsiveness that deals with the frequent unexpected changes in grades of service. The air medium and the mobility expected of ad-hoc multi-hop wireless networks requires careful and specialized higher layer protocols for congestion control and QoS support. In multimedia applications, this adaptability targets at maximizing the overall QoS delivered by the network and their functionality may be classified into (i) The transport functionality that decides the network parameters e.g. sending rate and (ii) The presentation functionality that decides the content that should fit the network parameters. In this paper we focus on (i). We are motivated by the fact that, in wireless it is particularly difficult to implement an accurate monitoring process (measurement) and embed it into a distributed strategy that efficiently controls the scarce network resources. Protocols developed for wired networks fail. The responsibility for flow control on the Internet is mainly left at the transport layer, allowing for a scalable design and a thin network layer. The transport peers perform some type of monitoring to their packets and apply sampling and estimation techniques to calculate desirable quantities e.g. trip times, path available bandwidth etc. Explicit help from lower layers is not allowed, as this would impair scalability, make deployment difficult and dramatically increase costs, especially router related. TCP for example, is using a flow's single packet loss as an indication of network congestion, presuming that the packet is dropped due some stressed buffer along its path. This however does not work in wireless networks, as packet losses due to external and internal interference are also frequent.

This inherent difficulty in distinguishing channel error from congestion related drops is affecting multimedia transports that use rate adaptation. Say, for example, a video transport is currently requesting a 256 Kbps from the codec and during the past monitoring interval it has lost 50% of the packets. If that loss occurred due to congestion related errors then it should start requesting less than 128Kbps from the codec. However if that loss is due to channel errors then the requesting rate should not drop. Instead the same or even more bandwidth should be requested to be used for increased Forward Error Correction bits (FEC) or retransmissions (ARQ). Therefore a successful add/drop policy should have a way of distinguishing between the two opposing possibilities.

An application scenario that has motivated this work is depicted in Figure 1, where multiple MANET clusters e.g. of Unmanned Air Vehicles, Unmanned Ground Vehicles form a large MANET to support collaborative efforts and on-demand real-time and near real time transmission of video with adaptive QoS¹.

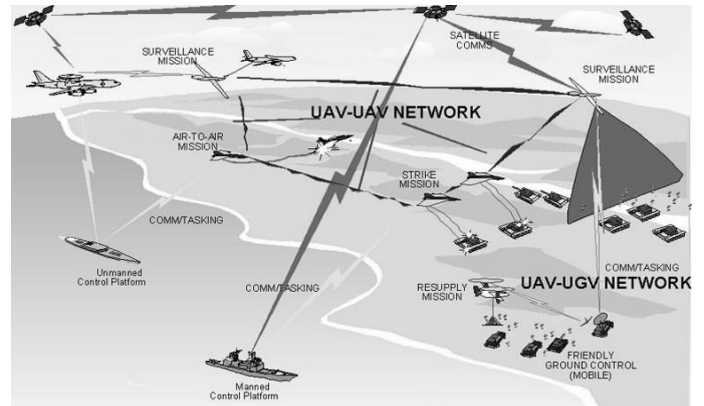


Figure 1. Video transmission through a UAV/UGV network in an Office of Naval Research scenario

B. REQUIREMENTS

Adaptive multimedia transports would ideally prefer to have accurate knowledge of the bandwidth available along their path, averaged over a small interval [Car97], [Pax97]. In a wireless setting a fairly accurate knowledge of available bandwidth has the additional advantage that it can be combined with loss rate information to help in distinguishing channel error versus congestion error. Let us define the available bandwidth over one link as the link bandwidth minus the used bandwidth, i.e. the un-utilized bandwidth. The path's available bandwidth then would be the minimum available bandwidth across all links in the end-to-end path. With this information at hand, the peers would be able to gradually adjust their rate so

¹ This work has been funded by the Office of Naval Research through Contract #N00014-01-C-0016

as to minimize the packets lost due to congestion, generally perform flow control and be TCP fair or friendly. Unfortunately, available bandwidth is very difficult to measure and filter using end-to-end techniques even in wired networks, because the observed samples follow a multi-modal distribution [Dov01].

In a nutshell, current end-to-end transport solutions for multimedia communication put to work in wireless networks are non-promising. Therefore, besides special development of end-to-end methods another option becomes particularly worth exploring, i.e. deploying network support for transports and applications. Let us call such architectures, network feedback architectures. These require special node support, possibly in both hardware and software. Each node measures its bandwidth and delay performance. This can be done fairly accurately because lower layers perform it, each knowing their own mechanisms. The values are then propagated using routing or other protocols. Eventually they reach the end hosts where they may be used by transports, applications or measurement based call admission algorithms. Such a setting has the advantage that it may overcome the aforementioned difficulties improving the overall performance and QoS. However, since each node requires special support, deploy-ability, scalability, interoperability and consequently cost are impaired. Note that per flow QoS is not required, just a per link QoS information estimation and a per routing table entry aggregate variable per QoS metric. But, how much would be the benefit of deploying such network support versus its cost in networks with wireless links?

In the next section we present our metrics and the QoS model, in section D we discuss the end-to-end architecture and add a different available bandwidth sampling technique in the options we are exploring. E describes the development of the architecture and measurement for lower layer feedback support, and F contains conclusive graphs of our experiments with the available choices. We conclude in G.

C. METRICS

The loss rate of the connection gives an incomplete idea of the QoS provided to that connection. The ultimate measure of performance in multimedia communications is the QoS delivered to the user. There has been extensive research in developing quantitative methods to evaluate QoS. These measures can be mathematical or perceptual, objective or subjective. In many applications, the Mean Square Error of a delivered image or sequence of images is expressed in terms of a *signal-to-noise ratio* (SNR) or PSNR (peak-to-peak SNR). These mathematical measures are not well correlated with human perception since they cannot take into account human vision properties and codec concealment techniques. The signal to weighted noise ratio, WSNR, is a PSNR metric that takes into account some human visual system properties through weighting luminance, power density and others against eye sensitivity (ITU-R 21451-2). Another measure has been proposed by the ITS called SHAT which has been optimized through subjective tests. In this study we use the Moving Picture Quality Metric (MPQM) metric [Van96]. The model is based on the properties of human vision. It considers visual masking techniques to effectively take into account error concealment techniques in the QoS result and directly targets

video rather than single image quality. It has been shown to behave consistently with human judgments [Van96]. Its quality rating is scaled from 1 to 5 as described in [Ard94]. The quality being Excellent (=5), Good, Fair, Poor and Bad (=1), and the Impairment being Imperceptible (=5), Perceptible, Slightly Annoying, Annoying and Very Annoying (=1).

For 5 Layers of 256,128,80,64,48 Kbps of high Motion, Frequent scene change clip, we have developed the following QoS model depicted in Figure 2. We see one line for each layer and their degradation to loss. This model is in accordance to the area of operation and techniques found in wireless [Vil98]. Due to space limitation more details should be found in [KazDs].

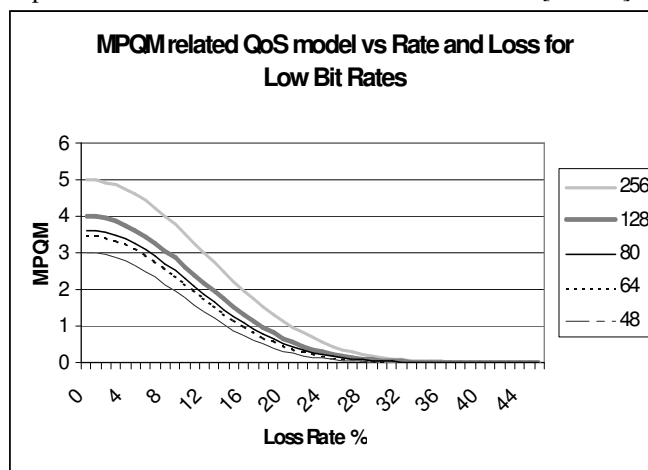


Figure 2 QoS MPQM related evaluation model used

D. THE END-TO-END RTP BASED ARCHITECTURE

This option is the most commonly used one, when congestion related multimedia adaptation is performed. It is implied in the RTP RFC. A QoS notification programming model is implemented to provide the server with the ability to monitor network conditions at the client end and react adequately when congestion occurs. Its activity consists of the following three parts: 1) it monitors the multimedia stream from the server to the receiver; 2) it measures QoS parameters of this stream, e.g. loss rate and jitter in the RTP case; and 3) it reports this QoS information to the application. In this paper we are experimenting with loss rate adaptation, which we consider a 'trial-and-error' adaptation and available bandwidth adaptation, a direct measurement method. The former uses two loss rate thresholds: a minimum below which an upgrade in quality is attempted and a maximum above which a downgrade is made. More details can be found in [Kaz99b]. In the direct measurement case we are estimating the available bandwidth and periodically try to fill a fraction of it. The estimation is either based on packet pairs and the 'bytes over time' model or alternatively the ab-probe model [KazT01].

E. THE LINK/NETWORK ASSISTED ARCHITECTURE

A less conventional, but promising as discussed earlier, approach is to employ lower layer explicit feedback mechanisms in place or in aid of end-to-end efforts. We name this, a "network feedback architecture". We use a single link, source-destination pair specific measurement, suitable for multi-hop networks. This measurement is then propagated and

aggregated throughout the network and made available to the sources using the routing messages. It is then used by transports at the sources through an API. The main disadvantage of lower level support of measurements is that nodes need to provide that support. Two or three extra fields per route contain the necessary information and distance vector, link state or on demand propagation is feasible.

The 802.11 standard allows for link performance measurements. It is a challenging task to measure the bandwidth available to applications in a multi-hop network due to the unique contention of each node and their dependence through the head of line blocking, hidden and exposed terminal situations. In order to produce a viable metric that takes into account the above we are measuring separately a source, destination specific achievable throughput and a per node utilization. We use the fragment acknowledgement and link-fail message used for unicast traffic in DCF to produce link-by-link (source, destination) pair throughput measurement. Link layer queue utilization measurements are then combined to produce an available bandwidth measurement, in a suitable way for a wireless multi-hop network (see also contention models in [Nad00]). Propagation of the available bandwidth values can be accomplished using piggybacking to routing messages or as a separate application. The former constrains the two mechanisms, with possibly different cost functions, to work on relevant time scales. In this paper we are using an on-demand event-driven measurement propagation and aggregation based on AODV [Per99], we call QM-AODV (see [KazDs]). Figure 3 shows an example scenario where an MREP, a special added Measurement support packet, is used. In our experiments with adaptation, the routing still operates without being affected by the propagation of these values, other than the increase in the size of messages, and calculates the minimum hop routes.

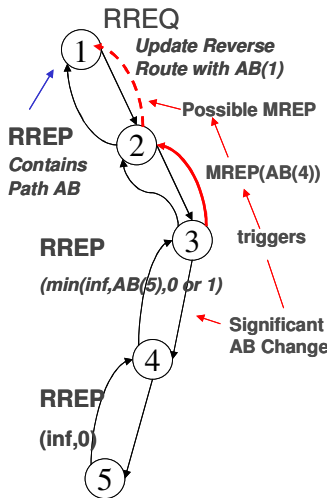


Figure 3. An example scenario where an MREP is used

The bandwidth reported as available to the applications, even considering a single hop connection, is a combination of the consumption by the node itself, the neighboring nodes, hidden and exposed terminals and head of line blocking dependencies. At the same time the measurement needs to be distributed and take place in one node without any exchange of information that add communication overhead. This can be

accomplished by timing specific 802.11 events that are affected by these situations. When a fragment is ready for transmission in the MAC layer we start to time. After that the 802.11 virtual carrier sense takes place, possible RTS/CTS exchanges as well as possible back-offs. Eventually the packet is transmitted and an acknowledgement or a link failure is returned. That defines the end of the fragment service time. In order to take into account the aforementioned characteristics we break down the measurement into two elements. One is the capacity of the link from the reference node to each neighbor. This is estimated by the bits over the service time. For example if the link bandwidth is 1Mbps the capacity of a neighboring (source, destination) pair affected by the distributed contention experienced might be 500Kbps (see also Figure 4). In a multi-hop network each node has a unique neighborhood and its traffic suffers unique competition. Therefore its capacity as we measure it here will change, taking into account the current environment through the carrier sense and retransmission delays. We then define our available bandwidth using a typical utilization measurement which has to do with the node itself. The utilization can be measured by calculating the idle time i.e. sent time of packet i - ACK receive time of packet $i-1$, over a window of packets. The queue utilization will be affected by head of line blocking and competing traffic that share the queue while the capacity will be affected by the traffic that causes contention, possibly due to hidden terminals but does not share the queue:

$$P(src, dest) = (1 - u) * Throughput(src, dest)$$

where u is the link queue utilization.

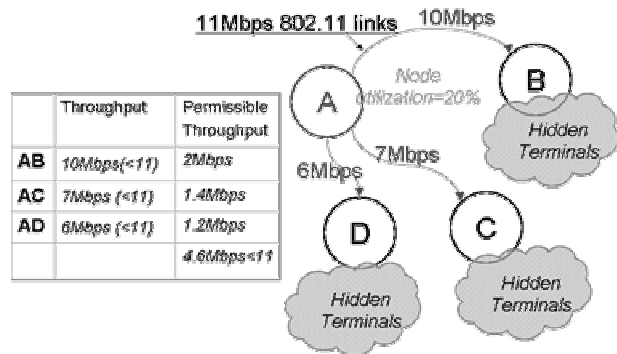


Figure 4. A numerical example of the throughput measurement, a source-destination capacity.

F. COMPARISONS

In this section we are developing a uniform environment under which we can test, by NS simulation, all the different options studied. We use the exact same network and the exact connections and mobility under different networks and all adaptation strategies. For each environment we report loss rates. As mentioned before loss rates are important because they can indicate whether and how much QoS may be improved. We then show QoS using our evaluation model introduced in Par. 2. Namely the curves shown in the graphs of this section correspond to:

- Non adaptive high layer/rate 180Kbps transmission
- Non adaptive low layer/rate 8Kbps transmission
- RTP 'Trial n Error' Loss Rate based Adaptation [Kaz99b]
- Packet Pair Available Bandwidth based Adaptation

- AB-probe sampling Available Bandwidth based Adaptation [KazDs]
- Network Feedback Adaptation

We run a large number of experiments with an increasingly larger number of connections. The connection source and destination points are decided at random but adhering to a hop count distribution. The probability of a connection being an h -hop connection is inversely proportional to h . The start and end times of the connection are also distributed probabilistically, first around 3 start and 3 end timelines and then uniformly within one second from the timeline chosen. In this way convergence of an adaptation is taken into account in the results. We use 7 pre-encoded layers at 180, 128, 88, 64, 32, 16 and 8 Kbps CBR codecs with paired packets. When mobility is introduced a random waypoint model is used. Specifically 75% of the 36 nodes randomly pick a destination every 5 seconds and move to it at the reported speed. We vary speeds from 2-55 km/h.

In Figure 5 we first note that the network feedback adaptation performs very well with respect to loss rates. It has almost the same loss as the low layer transmission. We also see that the AB-probe performs best of all the end-to-end solutions, handling up to 20 connections at a less than 20% loss. Traditional RTP adaptation works well over a single hop, showing some possibility for QoS improvement, depending on the codec. Now let us apply the QoS evaluation model we adopted to the loss rates as recorded every two seconds. A QoS value is calculated according to Figure 2 and then averaged over all connections and all time spans. This conclusive graph shows that with our high encoding complexity corresponding QoS model only Network Feedback strategies are capable of delivering better QoS across all network operation points. In fact it delivers improved QoS, or at worst the same QoS as the low layer does after more than 22 connections in the network. On the other hand end-to-end strategies quickly deteriorate in terms of the delivered QoS, delivering much lower perceptual quality than the non adaptive low layer scheme. The AB-probe adaptation is again the best among end-to-end techniques, improving QoS when connections are less than 10 and deteriorating after that.

Now let us look at the 802.11 multihop results. The conclusions are very similar to the one hop results. Now however, the network feedback strategy starts to digress from the lower layer non-adaptive transmission level of loss rates. Network Feedback delay, overhead and small accuracy error start to show with a large number of connections (close to the network capacity). When the network can handle 45 low rate connections without loss and 58 with less than 20% loss the network feedback case allows 30 connections without loss and 48 with less than 20% loss. Ab-probe performs best but starts performing worse than trial and error at the same point that network feedback loss is increased. It starts performing worse than RTP when loss rates are already very high at 45%. The same behavior was seen in the single hop case occurring at a 60% loss rate. RTP adaptation works well over a single hop, better than direct PP measurement.

In the mobility case, we show the 20 and 55 km/h graphs. Notice how RTP 'trial and error' adaptation is increasingly performing worse with increased mobility. The increased loss

rates are due to the inevitable mobility related losses and the initialization and convergence times of the filters.

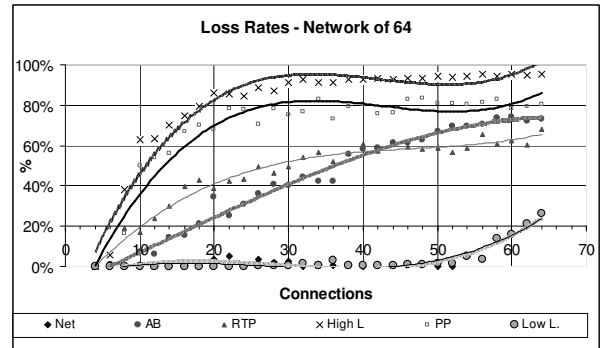


Figure 5 Loss Rates

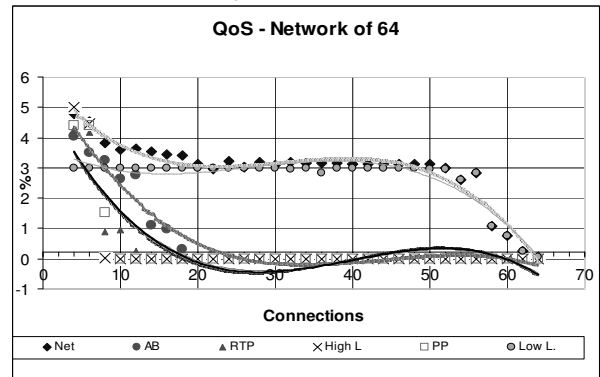


Figure 6. QoS, 802.11 single hop

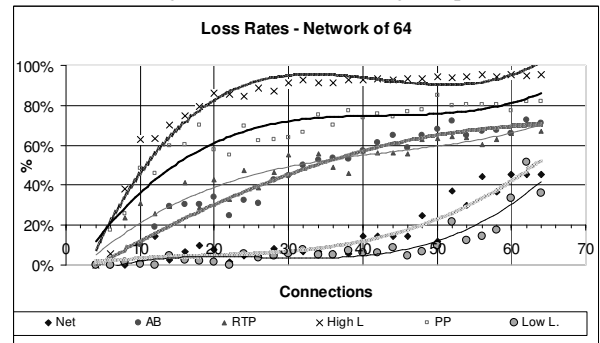


Figure 7 Loss rates on 802.11 multihop

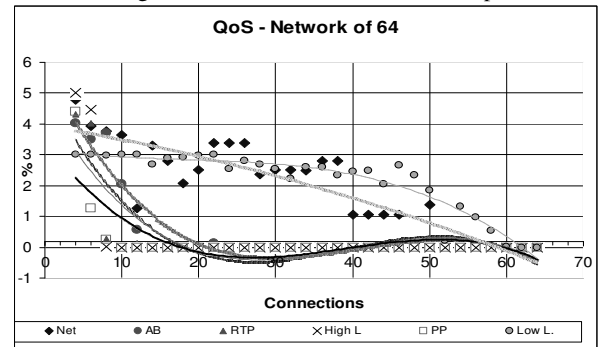


Figure 8 QoS on 802.11 multi-hop

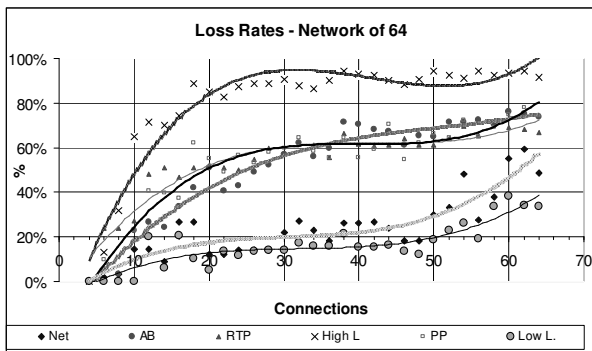


Figure 9 20km/h mobility

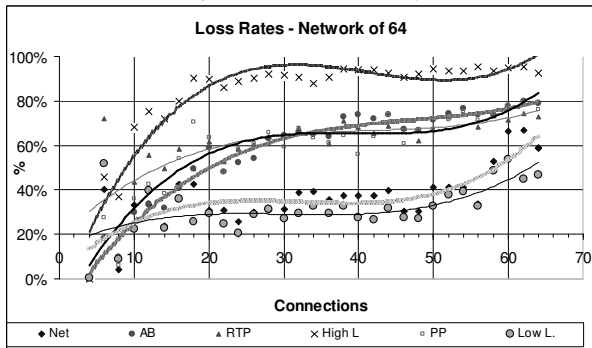


Figure 10 55km/h mobility

G. CONCLUSION

We first studied the RTP statistics based adaptation that has been used in wired networks and found that the overall QoS delivered to the user deteriorates fast with increasing number of connections. A much higher QoS would be delivered in that case if the connections did not adapt at all and were set at transmitting at low rates, given present day codec technology.

Direct end-to-end measurement approaches, as for example measurement of available bandwidth through packet pair dispersion become more attractive to deal with ‘trial and error’ slow convergence and response to network changes. We included in our study a new method of end-to-end available bandwidth measurement which was shown by some experimentation successful on heterogeneous networks. Using a high encoding complexity MPQM derived QoS model we see that with this available bandwidth measurement QoS is increased as compared to a non-adaptive low transmission QoS, when the number of connections is fairly small, 5 to 10 times smaller the low rate transmission case number (with the particular 7 rates we use).

This motivates the development of architectures that allow for lower layer feedback support. We develop link level support architecture in 802.11 that takes into account the particulars of the medium access and produces a very accurate measurement. We then use routing mechanisms to propagate those values at the sources for use in adaptation and/or call admission. This “network feedback” results in very low loss rates and increased QoS. Specifically the QoS delivered to the user is the best possible under the adaptation requirements, delivering at least the low rate transmission QoS in the vast majority of cases.

In summary we show that the benefits of network feedback support are large and necessary when one considers end-to-end congestion control performance in wireless multi-hop networks.

H. REFERENCES

- [Ard94] Ardito M., Barbero M., Stroppiana M. and Visca M. Compression and Quality. Proc. of the Intl. Workshop on HDTV 94, Brisbane, Australia, Oct 1994. Springer-Verlag.
- [Car97] R. L. Carter, M. E. Crovella, "Server selection using dynamic path characterization in wide-area networks," Proc. of IEEE INFOCOM '97, Kobe, Japan, pp.1014-1021, April 1997.
- [Dov01] Constantinos Dovrolis, Parameswaran Ramanathan, and David Moore. What do packet dispersion techniques measure? In Proc. of IEEE INFOCOM, April 2001.
- [Kaz99b] M. Kazantzidis, L. Wang, and M. Gerla, On Fairness and Efficiency of Adaptive Audio Application Layers for Multihop Wireless Networks IEEE MOMUC'99
- [KazDs] Manthos Kazantzidis, "Adaptive multimedia over Wireless IP Networks", UCLA Computer Science Dissertation 2002 (also <http://www.cs.ucla.edu/~kazantz>)
- [KazT01] M Kazantzidis, "How to measure available bandwidth on the Internet" – UCLA CS Technical Report #010032
- [Lai01] K. Lai, M. Baker, "Nettimer: A tool for measuring bottleneck link bandwidth", Proc. of the USENIX Symp. on Internet Technologies and Systems, San Francisco, CA, USA, March 2001.
- [Lai99] K. Lai, M. Baker, "Measuring bandwidth", Proc. of IEEE INFOCOM '99, New York, NY, USA, pp. 235-245, March 1999
- [Mas01] S. Mascolo, C. Casetti, M. Gerla, M. Yours truly., Sanadidi, R. Wang, "TCP Westwood: Bandwidth estimation for enhanced transport over wireless links" Proc. of Mobicom 2001, Rome, Italy, Jul. 2001.
- [Nad00] T. Nandagopal, T. Kim, X. Gao and V. Bharghavan, Achieving MAC Layer Fairness in Wireless Packet Networks. Mobicom 2000
- [Pax97] V. Paxson, Measurements and Analysis of End-to-End Internet Dynamics, Ph.D. Thesis, University of California, Berkeley, 1997.
- [Van96] Van den Branden Lambrecht C.J. and Verscheure O. Perceptual Quality Measure using a Spatio-Temporal Model of the Human Visual System. In Proc. of the SPIE, volume 2668, pages 450-461, San Jose, USA, February 1996.
- [Vil98] Niko Faerber, Bernd Girod, and John Villasenor "Extensions of ITU-T Recommendation of H.324 for Error-Resilient Video Transmission," IEEE Comm. Magazine, Vol. 36, page 120-128, June 1998