Power-Efficient Streaming for Mobile Terminals

Jari Korhonen*
Nokia Research Center
P.O. Box 100
33721 Tampere, Finland
jari.ta.korhonen@nokia.com

Ye Wang
National University of Singapore
3 Science Drive 2
Singapore 117543
wangye@comp.nus.edu.sg

ABSTRACT
Wireless Network Interface (WNI) is one of the most critical components for power efficiency in multimedia streaming to mobile devices. A common strategy to save power is to switch WNI to active mode only when network activity is expected. In streaming systems, this approach is problematic because data are typically received continuously. One solution is to transmit data packets as bursts, which leaves WNI more time between bursts in standby mode. However, that subjects bursty transmission in high peak rates, which leaves it prone to congestion. In this paper, we study theoretically and empirically the impact of burst length and peak transmission rate for observed packet loss and delay characteristics as well as potential energy savings in a Wireless Local Area Network (WLAN) environment. We outline and implement a test system with adaptive burst length to achieve improved trade-off between power efficiency and congestion tolerance.

Categories and Subject Descriptors
C.2.2 [**Computer-Communication Networks**]: Network Protocols – *applications, protocol architecture, protocol verification.*

General Terms
Algorithms, Performance, Experimentation.

Keywords
Multimedia Streaming. Power-Efficiency, Wireless Networking, WLAN

1. INTRODUCTION
The rapid evolution of mobile computing and telecommunications is turning mobile phones into fully equipped entertainment centers capable of reproducing live video and music. As multimedia streaming gains popularity among mobile users, power management becomes increasingly important for streaming applications.

NOSSDAV'05, June 13–14, 2005, Stevenson, Washington, USA.

Power efficiency in wireless multimedia has been studied actively during the recent years [1, 2]. Research in this field can be divided roughly in two categories: power-aware multimedia decoding [3, 4] and power-efficient radio communications [5, 11, 12]. It is well known from the literature that WNI is typically the most critical part for power consumption in mobile terminals [5]. This is why standards for wireless data communications usually present a power save mode allowing WNI to go into sleep state during the inactive periods. During the sleep state WNI is not able to receive any data, but the power consumption is much lower than in active state.

In IEEE 802.11 WLAN, for example, the receiver in power save mode wakes up periodically for a short time. The periods are called beacons. If the base station wants to transmit data to the mobile terminal, it notifies the receiver during a beacon so that the receiver knows to stay in active mode when the transmission comes through.

However, streaming applications typically transmit a constant stream of packets arriving periodically. The traditional approach for power saving is not effective for this kind of applications, because the gaps between packets are short, and there are virtually always data packets waiting for transmission at the base station. That means the receiver has no time to sleep between the active periods.

If the streaming application tolerates reasonable buffering delay, the problem can be solved by reshaping the traffic so that packets are sent as bursts instead of keeping constant intervals between packets. This method allocates longer gaps between packet bursts. The receiver can easily switch to the sleep mode during inactivity.

In general, traffic burstiness is an undesired feature in packet-switched networking. Clustered packet arrivals may cause congestion in routers or overflows in transmitter buffers, leading to packet losses and decreasing overall network performance. In addition, bursty transmission does not guarantee that packets are still clustered close together when they reach the receiver. In fact, heavily clustered packet arrivals in network devices may increase random variation in transport delay, jitter, increasing dispersion of the clustered packets. This makes it even more difficult to predict the packet arrival times at the receiver. For this reason, many researchers have proposed avoiding bursty transmission, and reshaping traffic as close to the last link as possible instead.

*) This work was done during appointment at the School of Computing, National University of Singapore.

The strategy was first proposed by Chandra and Vahdat in [10]. They used an application-specific proxy to reshape the packet traffic of well-known multimedia formats. Another proxy-assisted method was proposed by Shenoy and Radkov in [9]. In their proposal the proxy provides additional information to the receiver for use in improving the prediction of packet arrival times. Acquaviva et. al. have also studied different alternatives for power-efficient scheduling based on traffic shaping [12].

In [5], Zhu and Cao presented a scheme for reshaping streaming media traffic for receivers using a shared radio link. They highlighted the importance of transition delay between active and sleep states. The energy consumption during transition is the same as in active mode, and therefore it is not efficient to switch WNI to sleep mode if the length of the inactive period is supposed to be shorter than the transition from on to off and vice versa.

In practice, it is often not feasible or economical to involve an application aware base station. We propose a suboptimal solution employing a layered multimedia coding scheme with customized packet scheduler providing bursty traffic with decreasing priority order of packets in each burst. The receiver may maintain stable power efficiency by sacrificing some of the enhancement layer data in case of highly scattered packet bursts.

In this paper, we analyze the delay properties of packets in bursty transmission mode. We show that although long bursts theoretically improve power efficiency, they also lead to greater dispersion of end-to-end transport delays and more congestion-related packet losses. To solve this dilemma, we propose a streaming system using adaptive burst length. Experimental results show that our approach provides a good trade-off between power efficiency and robustness against network congestion.

2. THEORETICAL BACKGROUND

Power-aware packet scheduler should trade-off between the advantages (good power saving ratio) and disadvantages (increased probability of congestion) of bursty packet transmission. This can be achieved by selecting the burst length and packet transmission intervals appropriately.

If the peak transmission rate during bursts exceeds the maximum available throughput of the slowest link between the sender and the receiver, buffers at the edge of the bottleneck would be filled up quickly. Obviously, this leads to increased delay and even loss of the last packets in the burst. Wireless access links are usually (but not always) the bottleneck. Unfortunately, throughput is not usually known by end systems and it is often dynamic by nature. This is why the application should be able to approximate the throughput adaptively via measurements.

Packet prioritization may also be considered in the scheduling policy. Typically, a multimedia stream consists of frames of different priority. In addition, blocks of different perceptual significance could be extracted from video or audio frames and interleaved among different packets [6, 7, 8]. In this kind of schemes losses of low priority packets can be easily concealed without significant loss of reproduction quality. The critical data should be delivered as reliably as possible, but higher packet loss rate is acceptable for the lower priority packets.

If the network conditions are good and the average transmission rate is well below the maximum network throughput, even relatively long bursts can be considered. In this case packet loss rate may get high during occasional congestion periods and especially in the end of a burst, but this is not fatal as mostly just the low priority packets are concerned. This is the motivation to allocate the high priority packets in the beginning of each burst.

The Relative One-way Transport Times (ROTT) of packets can be computed if the receiver knows the corresponding time of transmission for each packet from a timestamp with known resolution included in each packet. This information can be used to predict the arrival time of the first packet in each burst.

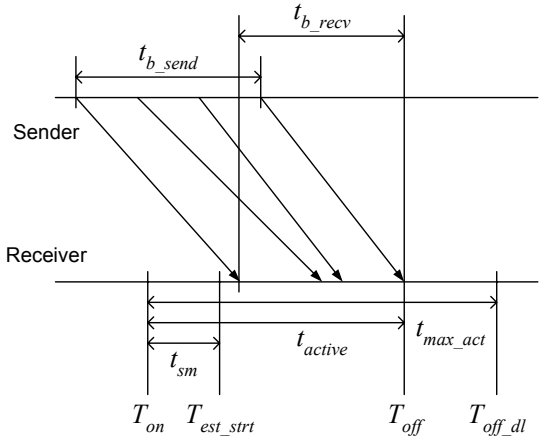

Figure 1. Timing for bursty communications.

The sequence diagram in Figure 1 summarizes the notation of timing in the receiver system. The time between the first packet and the last is the burst length at the sender, t_{b_send}. The predicted earliest possible arrival time for the first packet of the burst is T_{est_strt}. It is estimated from the arrival time history data collected so far. There is a safety margin t_{sm} between T_{on}, when the receiver switches the WNI into active mode, and T_{est_strt}. A safety margin is needed for two reasons: there is an off-on delay when the WNI state changes, and under changing network conditions it is possible that the first packet arrives before T_{est_strt}.

When all packets within a specific burst have been received, the system can switch WNI back to sleep mode until the next burst is supposed to arrive. This time is indicated as T_{off}. The active period is the time t_{active} between T_{on} and T_{off}. However, sometimes packet(s) may get lost or delayed so much that it is reasonable to stop waiting for them. This is why we may want to define a deadline T_{off_dl}. WNI switches to sleep mode at T_{off_dl} even if there is a packet or more still under way to be received.

Obviously, in an ideal system, the active period length t_{active} would be as close as possible to the actual observed burst length t_{b_recv} at the receiver system. In addition, t_{b_recv} should be as short as possible. On the other hand, T_{on} and T_{off_dl} should be selected rather conservatively to avoid losing packets during the sleep state.

Given the requested average media transmission rate B, the number of packets in each burst n and payload length in each packet P, the interval between the start of two adjacent bursts

should be $t_{cycle} = nP/B$. The average sleep time t_{sleep} of WNI during each cycle can then be computed with (1):

$$t_{sleep} = \frac{nP}{B} - (t_{active} + t_{on\_off}) \quad (1)$$

In (1), t_{on_off} is the transition delays introduced between active mode and sleep mode. To optimize power efficiency, the proportion of sleep time in each cycle should be maximized. Therefore, we should select n and the gaps between packet transmissions during a cycle t_{int} so that equation (2) reaches its minimum value.

$$\frac{t_{sleep}}{t_{cycle}} = 1 - \frac{B(t_{active} + t_{on\_off})}{nP} \quad (2)$$

Unfortunately, it is not known exactly how changing n and t_{int} impacts t_{active}, as that depends on the link bandwidth and congestion conditions. Intuitively, we can assume that a large value of t_{int} leads to a large t_{b_send}, and thus to a large t_{active} also. On the other hand, very small values of t_{int} would increase t_{active} due to congestion delays caused by heavy peak traffic. The same applies to burst length n; we can see directly from (2) that long burst increase sleep time, but large n could lead to longer traffic peaks prone to congestion and therefore may increase t_{active}.

If we know the length of the transition between sleep and active states, we can solve the theoretical relative power saving with different parameters. According to the literature [5], the on-off transition time is typically in order of tens of milliseconds. In the following example, we assume that the transition period t_{on_off} is 20 ms (10 ms from sleep mode to active mode and another 10 ms back to sleep mode). In addition, we assume that t_{b_send}, t_{b_recv} and t_{active} are equal.

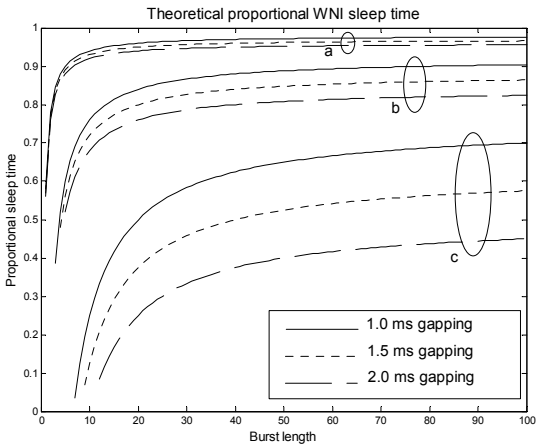

Figure 2. Proportional maximum theoretical power savings.

In Figure 2, the theoretical proportion of time spent in sleep mode with different parameter settings is illustrated. We have derived the curves by applying the equation (2) with the abovementioned assumptions. Set (a) shows the power saving for a 128 kbit/s stream with a packet payload size of 800 bytes, resulting in 20 packets per second. Set (b) shows similar results with 512 kbit/s, 800 bytes per packet, 80 packets per second, and Set (c) with 2 Mbit/s, 1000 bytes per packet, 250 packets per second, correspondingly. The curves have been drawn using three different transmit intervals, 1.0 ms, 1.5 ms and 2.0 ms. These are realistic defaults in a WLAN environment.

As the figure shows, a better power saving ratio can be achieved with a lower bitrate, shorter gapping and longer burst length. On the other hand, these parameters are bound by user requirements and network restrictions. Bitrate is directly related to quality of audio or video encoding. Transmission interval cannot be shortened below the threshold derived from peak network capacity. Long bursts are not only harmful for the overall network performance, but also require larger buffers and longer buffering delays.

To verify the theoretical analysis we have conducted practical experiments to evaluate the packet loss and delay characteristics in a real WLAN environment. Our test server was a regular desktop PC connected to our campus network. The test client is ran on a laptop computer that was connected to the campus network via the WLAN access points (APs). Standard application programming interfaces were used for UDP communications.

After preliminary tests, we selected four different burst lengths (10, 25, 50 or 100 packets) and four different packet transmission intervals (1.0 ms, 1.5 ms, 2.0 ms, and 2.5 ms) to be tested at UDP payload size of 1000 bytes and overall transmission rate of 2 Mbit/s. With the given parameters there are 250 packets transmitted per second. Gapping of 1.0 ms results in approximately the same peak transmission rate as the theoretical maximum bitrate of IEEE 802.11b standard, which is 11 Mbit/s. We selected a high bitrate to provide greater rigor in our experiments.

Each test case was repeated at least three times under both congested and non-congested conditions. Contending background traffic was generated by using three other laptops in addition to the test client, each receiving a 512 kbit/s video stream from the external servers in the Internet. Each laptop was located physically near each other; thus they shared a common WLAN AP.

The results of the tests are in exact numeric values as well as more general observations. In general, the impact of burst length seems to be essentially more important than the impact of transmission interval. With longer bursts, the stream is more likely to be stalled due to occasional heavy packet losses. No similar difference was found with different gap lengths when burst length was constant.

Figures 3 and 4 show the typical cumulative distributions of relative transport times with congestion and without. Arrival times are relative to the shortest observed difference between the receiver time and sender timestamp of a first packet within a burst. In both cases, the burst length is 25 packets and transmission interval 1.0 ms. Numbers denote the packet order number within a burst. To avoid cluttering, only six curves are given in each figure. Figure 3 shows the cumulative distributions when there is no contending traffic, and Figure 4 when there is congestion present, respectively. Comparing the figures, the difference can be seen clearly. In both cases, the last packets in

each burst suffer from higher dispersion in relative arrival times. Different test cases showed similar behavior.

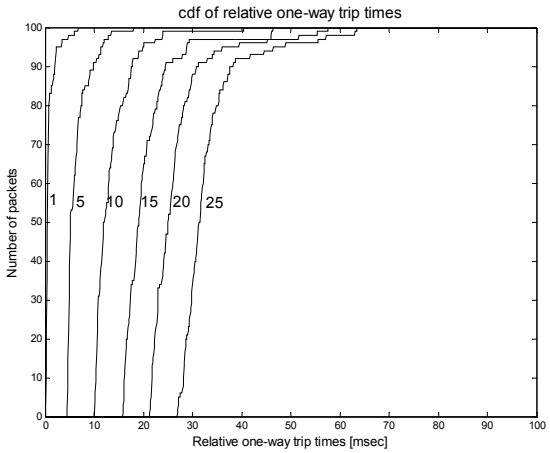

Figure 3. Cumulative distribution of relative arrival times for bursts of 25 packets and 1.0 ms gaps with no congestion.

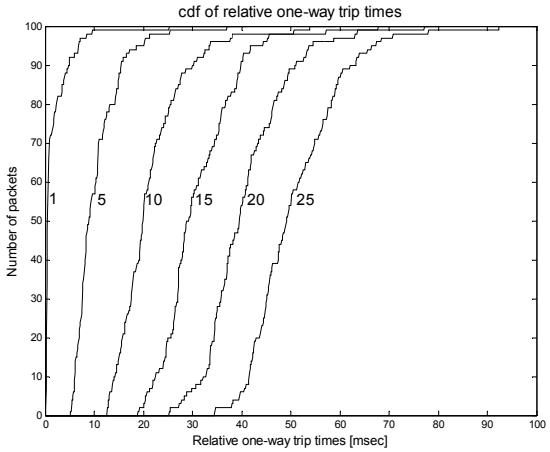

Figure 4. Cumulative distribution of relative arrival times for bursts of 25 packets and 1.0 ms gaps with congestion.

Packet loss is another important factor to be evaluated. According to the hypothesis, packet loss would be higher at the end of each error burst. Although some test runs show different behavior, the test results support the hypothesis in general. Especially at longer bursts the increasing tendency for packet losses across the sequence number is clearly visible.

3. ADAPTIVE STREAMING SYSTEM

3.1 System Design Principles

A power-aware adaptive streaming system requires a system-aware packetizer, a decision module at the receiver to select the appropriate burst lengths and packet transmission intervals, as well as a control protocol to convey feedback messages from the receiver to the sender. The control protocol may be implemented using application-specific Real-Time Control Protocol (RTCP) messages or a proprietary feedback system.

According to the experimental results, burst length is more important than gapping. Proper gap length can be selected by trying out different gaps to find the peak transmission bitrate threshold and adding a reasonable margin. Burst length selection is the main concern in the adaptation algorithms.

The most straightforward method to implement the required functionality is to divide the streaming process in two phases. In the first phase the system trials for an optimal transmission interval. In this phase the sleep mode is not used. When a satisfactory gap length has been found, the system switches to the power save mode.

In the bursty transmission mode, the receiver sends feedback information to the sender. The sender is responsible for decreasing or increasing burst length to cope with the prevailing conditions. As increasing packet loss is the most essential indicator of congestion, the system must react quickly to packet losses by shortening burst length. Increased jitter may be taken as a hint of congestion as well. In case of decreasing jitter or packet loss, the sender can try to increase burst length.

When a priority-based packetization scheme is used as described in Section 2, burst length cannot be selected arbitrarily. This is because the decreasing order of packets according to their priority should be maintained. Figure 5 shows the basic principle of generating priority ordered bursts of different lengths out of the original sequence of a fixed number of packets. Packets with the same priority are spread evenly among each burst via simple interleaving.

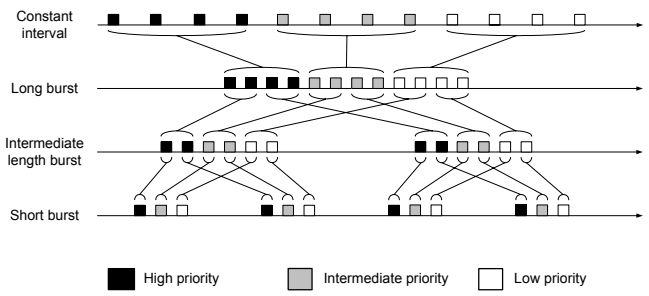

Figure 5. Generation of packet bursts with packets of different priorities.

The proposed method of rearranging packets does not require any kind of re-encoding or transcoding. The algorithm for burst length optimization is rather straightforward. The average transmission rate is kept the same, only the burst length changes. As rate adaptation requires changing the bitrate of the original encoded bitstream, we assume that rate adaptation is implemented separately and omit it from our study.

Burst length adaptation is based on two fundamental metrics: packet loss and jitter. Several clustered packet losses indicate severe congestion and the system must turn off the bursty transmission mode immediately. Changes in jitter may indicate more subtle changes in network conditions. This is why the system should decrease the burst length if jitter increases significantly or vice versa when the jitter decreases, respectively.

Jitter j is computed using function (3), where D_1 and D_n are the differences between the receiving time and the timestamp for the first and the last (n:th) packet in the current burst, respectively. Because jitter is higher for the later packets within a burst, it is reasonable to use the first and the last packet for jitter computations to approximate average jitter. Smoothing factor α (0..1) defines the sensitivity of the function and n is the burst length.

$$j_i = \alpha j_{i-1} + (1-\alpha)\frac{(|D_n - D_1| - j_{i-1})}{n} \qquad (3)$$

Using different values for α, it is possible to distinguish between long term jitter and short term jitter. When short term jitter gets substantially higher or lower than long term jitter, it is indication of a change in network conditions. The system should then react by changing burst length.

The feedback mechanism is straightforward. The receiver waits for a burst, updates short term and long term jitter, and sends ACK messages to the sender regularly. If a substantial amount of packets have not been received within the expected time limit, receiver does not send a feedback message. Missing ACK tells the sender to switch to non-bursty transmission mode immediately. Sender makes the decision about burst length after every burst when the feedback message has been received, according to the statistics included in the message.

3.2 Test Implementation

We implemented a test system to try out the proposed adaptive streaming scheme. Adapting the concepts from the audio streaming systems presented in [6, 8], we grouped 64 packets into a basic cycle. Each cycle consisted of 8 high priority packets, 8 intermediate priority packets and 48 low priority packets. In the test implementation four different burst length options are used: 64 packets, 32 packets, 16 packets and 8 packets. The decision about reducing or increasing the burst length was made as explained above.

We tested our proposed streaming system by comparing the potential energy saving with that when using adaptive burst length and fixed burst length in a real WLAN environment. We assumed that the total transition delays were 20 ms per each cycle. Figure 6 shows the proportion of time spent in the sleep mode. Because power is saved in the sleep mode, this value describes the relative power efficiency in the WNI. The figure shows the results of three test runs for four different transmission modes: adaptive burst length and fixed burst length of 16, 32 or 64 packets. In ideal circumstances, the performance of an adaptive scheme and constant burst length of 64 packets would be close to each other, because the adaptive system would use the maximum burst length most of the time. This is why we introduced some disturbance by running other streaming applications in the same WLAN during each test run.

As expected, longer burst length provides higher power saving in most cases. The adaptive scheme outperformed the fixed bursts of 16, 32 and 64 packets in all test runs. However, the performance depends highly on the network conditions. Because long bursts are especially vulnerable to congestion-based jitter, bursts of 32 packets performs better than bursts of 64 packets when the congestion is most severe (the first test run).

Figure 7 shows the average packet loss rate from the three test runs, computed separately for the packets of different priority class. With small burst lengths packet loss is lower in general. Packet losses are distributed most smoothly among low and high priority packets when burst length is smallest (16 packets) or the adaptive scheme is used. The adaptive scheme shows a clear advantage against long bursts of fixed length.

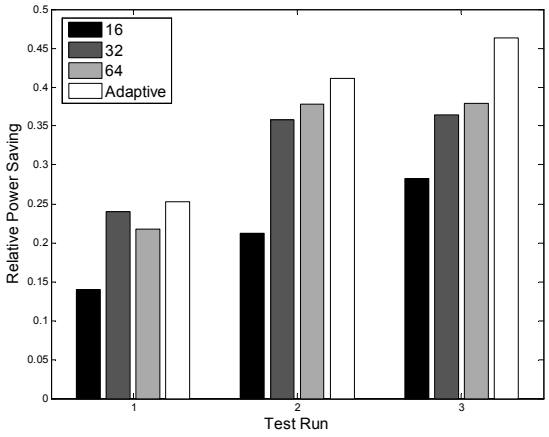

Figure 6. Power saving in different streaming modes.

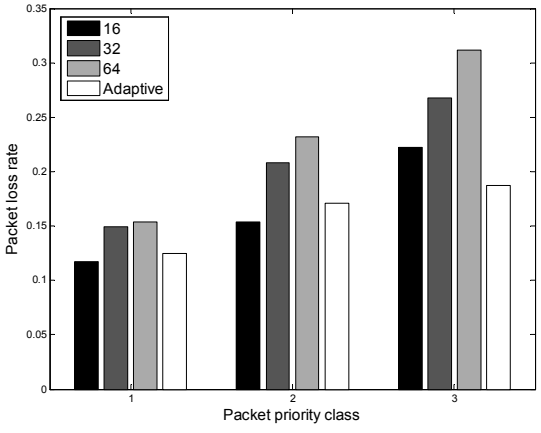

Figure 7. Packet loss rates for different priority classes (average from three test runs).

In general, the experimental results support the hypothesis of long bursts providing generally good power efficiency, but being less robust against congestion related increase in jitter and packet loss. In addition, long bursts allow better differentiation in packet

prioritization, because packet losses would be more clearly concentrated at the end of the burst.

4. DISCUSSION

In this paper, we have focused on rate shaping in either the streaming server itself or a local proxy. In principle, it is possible to implement a rate shaping mechanism in the wireless base station, such as a WLAN access point. This would solve the congestion problem in the wireline part of the network, as traffic would be bursty only in the wireless access link, which is the last link between the sender and the receiver. Another advantage of this arrangement would be the possibility of deploying a power-aware multiple access scheme to accomplish the proposed rate-shaping mechanism.

In theory, performance of the system could be substantially improved also by adding supportive functions in the wireless hardware. For example, power-aware WLAN access point could delay delivery of first packets if it knows that the mobile terminal is still in the sleep mode. However, there are several practical difficulties with implementing an application layer traffic shaping system in a network element such as WLAN AP. Most importantly, APs must be made application aware because the packet reordering procedure requires relatively detailed knowledge of the packetization principles at the end system.

In our work we have mostly omitted the role of resource allocation schemes for wireless networking, such as Carrier Sense Multiple Access / Collision Avoidance (CSMA/CA) used in IEEE 802.11 WLAN. These mechanisms work typically in a short timeframe and they may confuse the bursty transmission if the burst length is short in proportion to the time slots allocated for each user. This is one more reason why burst length must be reasonable long in order to gain any benefit from the scheme. Optimally, the number of other users sharing the wireless link is limited, or at least, they are not all using same kind of bursty transmission mechanism.

5. CONCLUSIONS

In wireless multimedia streaming, power can be saved if data packets are transmitted as bursts, which allows the receiver to switch to energy-saving sleep mode between bursts. However, traffic peaks caused by bursty transmission can be harmful for the network performance. In this paper we have studied theoretically and experimentally the impact of packet transmission interval and burst length on observed network performance characteristics, such as transport delay, jitter, and packet loss rate. Long bursts provide the best power efficiency, but they also increase the risk of network performance problems in a shared link.

As a solution to this dilemma, we have proposed an adaptive energy-saving streaming mechanism that adjusts the burst length to the prevailing congestion conditions. Our experiments with a test application show that this approach provides good trade-off between good power efficiency (average burst length and low dispersion of bursts) and congestion avoidance (packet loss and network problems observed by other users).

6. REFERENCES

[1] Benini, B., and De Micheli, G. System-Level Power Optimization: Techniques and Tools. *ACM Transactions on Design of Automatic Electronic Systems, 5*, 2 (Apr. 2000), 115-192.

[2] Jones, C. E., Sivalingam, K. M. , Agrawal, P. , and Chen J. C. A Survey of Energy Efficient Network Protocols for Wireless Networks. *Wireless Networks, 7*, 4 (Aug. 2001), 343-358.

[3] Pouwelse, J., Langendoen, K., and Sips, H. Application Directed Voltage Scaling. *IEEE Transactions on VLSI Systems, vol. 11*, 5 (Oct. 2003), 812-826.

[4] Choi, K., Dantu, K., Cheng, W-C., and Pedram, M. Frame-based dynamic voltage and frequency scaling for a MPEG decoder. In *Proceedings of the IEEE/ACM International Conference on Computer Aided Design (ICCAD '02)*, (San Jose, California, November 10-14, 2002), 732-737.

[5] Zhu, H., and Cao, G. A Power-Aware and QoS-Aware Service Model on Wireless Networks. In Proceedings of the *IEEE Infocom '04*, (Hong Kong, March 7-11, 2004), 1393-1403.

[6] Korhonen, J. Robust Audio Streaming over Lossy Packet-Switched Networks. In *Proceedings of the International Conference on Information Networking (ICOIN '03)*, (Jeju Island, South Korea, February 12-14, 2003), 1343-1352.

[7] Varsa, V., and Karczewicz, M. Slice Interleaving in Compressed Video Packetization. *In Proceedings of the Packet Video Workshop*, (Forte Village, Italy, May 1-2, 2000).

[8] Wang, Y., Huang, W., Korhonen, J. A Framework for Robust and Scalable Audio Streaming. In Proceedings of the *ACM Multimedia '04*, (New York, USA, October 10-16, 2004), 144-151.

[9] Shenoy, P., and Radkov, P. Proxy-Assisted Power-Friendly Streaming to Mobile Devices. In *Proceedings of the Conference on Multimedia Communications and Networking (MMCN '03)*, (Santa Clara, California, January 29-31, 2003).

[10] Chandra, S., and Vahdat, A. Application-Specific Network Management for Energy-Aware Streaming of Popular Multimedia Formats. *In Proceedings of USENIX '02*, (Monterey, California, June 10-15, 2002).

[11] Havinga, P., and Smit, G. Energy-Efficient Wireless Networking for Multimedia Applications. *Wireless Communications and Mobile Computing, Wiley, 1*, 2 (Apr-Jun 2001), 165-184.

[12] Acquaviva A., Lattanzi, E., Bogliolo, A. Design and Simulation of Power-Aware Scheduling Strategies of Streaming Data in Wireless LANs. In *Proceedings of the International Workshop on Modeling Analysis and Simulation of Wireless and Mobile Systems (MSWIM '04)*, (Venice, Italy, October 4-6, 2004), 39-46.