Energy-Efficient Multicast with Deadlines in Wireless Networks via Lazy Rate Scheduling

Yanan Bao, Xiaolei Wang, Sheng Zhou, Zhisheng Niu
Tsinghua National Laboratory for Information Science and Technology, Dept. of Electronic Engineering, Tsinghua Univ., Beijing, 100084, P.R. China
{byn10, wang-xl09}@mails.tsinghua.edu.cn, {sheng.zhou, niuzhs}@tsinghua.edu.cn

Abstract—In this paper, we consider the problem of providing QoS (Quality of Service) guaranteed data download services in wireless networks with high energy-efficiency. To guarantee the QoS, a transmission deadline is set for each request, before which the whole data should be received by the user. To attain high energy-efficiency, multicast technique is adopted. Traditionally, the multicast rate is constant and independent of the number of users served in one session. However, we show that when we change the transmission rate dynamically according to current state, such as user numbers, more transmission energy can be saved. We design the optimal rate scheduling algorithm for the offline case when non-causal information about user arrivals is available. We show that the optimal rate scheduling scheme can save energy by 15%. For the online scheduling problem when only causal information is available, we propose several heuristic lazy transmission schemes based on DP (Dynamic Programming) results. Simulation results show that with these heuristic schemes, up to 10% transmission energy can be saved.

I. INTRODUCTION

With the rapid development of mobile communication networks, voice traffic is gradually giving way to data traffic with multimedia contents. Statistics from Cisco shows that the dominant traffic nowadays has been mobile video [1]. These video traffic usually has QoS(Quality of Service) metrics, like delay and stream rate, etc. However, in today’s mobile networks, most transmission schemes are best-effort, and thus consume great amount of energy when they guarantee QoS for data services. Some studies show that ICT (Information and Communication Technology) industry is responsible for 2%-2.5% of the global carbon emissions [2] and the percentage is expected to grow further. Therefore, it is urgent to improve energy-efficiency of the wireless networks to support QoS guaranteed multicast services.

For data services, the user preferences are aptly converged to some popular contents. Also the prosperity of SNS (Social Networking Services) helps to improve the convergence. For instance, 10% of news items account for 70% of the total traffic of the www.sina.com website [3]. Moreover, some statistics show that in a network of only 200 users, the most popular YouTube videos are requested 6 times in just 10 minutes [4]. For these popular contents, multicast is a good choice to reduce the redundancy caused by multiple unicasts. In fact, the 3rd Generation Partnership Project (3GPP) has proposed multimedia broadcast and multicast services (MBMS) [5] to save network resources in cellular networks.

There have been researches [3], [6], [7] in literature focusing on reducing the transmission energy consumption in wireless systems. For instance, Ref. [6] utilizes techniques, like power control, channel selection and rate splitting, to reduce power consumption in MBMS. Ref. [3] proposes to use relay stations (RSs) to cache popular contents, such that they can serve the nearby users that possibly request cached contents in the near future. Ref. [7] proposes a lazy packet scheduling scheme to minimize transmission energy subject to a deadline for all packets in a unicast wireless system. However, little work deals with the problem of how to efficiently schedule transmission rate while providing QoS guaranteed services for multicast.

In this paper, we minimize energy consumption through rate scheduling for wireless multicast while guaranteeing the QoS of each user. For each user who requests a data download we set a deadline, before which the complete data should be reliably received by the user. We find that transmission rate should adapt to current state (e.g. the number of users, the data received, the remaining time of each user before its deadline, etc.) for energy saving, rather than keep constant during the multicast session. The main focus of this paper is to find the optimal rate scheduling algorithm. We derive the optimal solution of the offline scheduling problem, in which non-causal information about user arrivals is assumed. For the online scheduling problem, dynamic programming (DP) is adopted to obtain the solution for a simplified case with only two users. Based on these results, we propose several heuristic lazy rate scheduling schemes. Simulations are performed under different traffic models to evaluate the energy-efficiency of our schemes.

The remainder of this paper is organized as follows. In Section II, we depict the system model. The offline scheduling problem is formulated and solved in Section III, and then we proceed to the online scheduling problem in Section IV. Finally, we conclude our work.

II. THE SYSTEM MODEL AND PROBLEM FORMULATION

A. Energy consumption model

For simplicity, we assume the wireless channel from the transmitter to each receiver is an identical and time-invariant Additive White Gaussian Noise (AWGN) channel with bandwidth W and noise power density N_0. The channel fading is not considered and the channel gain from the transmitter to
each receiver is assumed to be 1. Then, the channel capacity is:

\[ C = W \log_2 \left( 1 + \frac{P}{N_0 W} \right) \]  

(1)

where \( P \) denotes the transmission power.

Considering the overhead of channel coding, information can be reliably distributed at a rate \( R = \alpha C \), where \( \alpha \in (0, 1) \). Therefore, to distribute data at a rate of \( R \), the transmission power should be

\[ P(R) = N_0 W \left( 2^{\frac{R}{W}} - 1 \right). \]  

(2)

Given a segment of data with \( L \) bits, the transmission duration will be \( \frac{L}{R} \) at a transmission rate \( R \). Thus, the energy consumption to distribute the data is

\[ E(R) = \frac{N_0 W L}{R} \left( 2^{\frac{R}{W}} - 1 \right). \]  

(3)

It is easy to prove that \( E(R) \) is monotonically increasing and strictly convex in \( R \) and, when \( R \) approaches to zero, the energy required to distribute the data is minimized, i.e., \( E(0) \) approaches \( \frac{N_0}{2} \ln 2 \). Note there is a tradeoff between energy consumption and transmission delay [8], [9], [10]. Here we set a transmission deadline \( T \) for each request as the QoS requirement.

\section*{B. Traffic model}

In terms of the traffic, two types of user arrival models are considered: one is Poisson process with an average arrival rate \( \lambda \), and the other is a 2-state Markov modulated Poisson process (MMPP) with parameters \((\lambda_1, \lambda_2, s_1, s_2)\) [11]. The MMPP2 is introduced to represent traffic arrival with positive temporal correlation. The parameters \( s_1^{-1}, s_2^{-1} \) denote the mean sojourn times of the two states, and \( \lambda_1, \lambda_2 \) denote the average arrival rates in the two states, respectively. Without loss of generality, we assume \( \lambda_1 \geq \lambda_2 \), indicating that state 1 corresponds to the busy mode and state 2 corresponds to the idle mode. So the average arrival rate \( \lambda \) in this model equals \( \frac{\lambda_1 s_2 + \lambda_2 s_1}{s_1 + s_2} \).

\section*{C. Service model}

We focus on the data download service (e.g. software update, movie download, news distribution, and so forth). The main characteristic of these services is that they do not require playout while downloading, but demand the accuracy and integrity of the received data. For simplicity, we only consider one segment of data with \( L \) bits that is required by all users. Once a user requests the data, the server setups a multicast session (The formal definition of multicast session in this scenario will be given in Section III) and starts to transmit the data. Any other user who requires the same data can join in the multicast session anytime before the end of the session. Different from traditional multicast session, the length of each session is variable and depends on the user arrivals. To guarantee the QoS of each user, a corresponding deadline is set, which means the data of length \( L \) should be reliably transmitted within a period of length \( T \). We use Fountain Codes [12], [13] in application layer to facilitate multicast.

The benefit is that the data transmission is not interrupted by users’ arrivals and departures.

In the traditional multicast system, the transmission rate is constant in a multicast session. For instance, with a constant transmission rate \( R = \frac{C}{T} \), the transmitter can guarantee that all user can receive sufficient data right by their deadlines. However, we are going to show that some benefit, such as energy-efficiency, can be obtained by changing the transmission rate dynamically in a multicast session. Intuitively, it is more energy-efficient to increase the rate when there are more users and decrease it when there are less. The goal of this paper is to find the optimal rate scheduling algorithm.

Suppose there are \( U \) users and the \( i \)-th user arrives at time \( t_i \). The corresponding departure time of the \( i \)-th user is denoted by \( t_i' \). Due to the deadline constraint, the inequality \( t_i' \leq t_i' + T \) should be met. Our target is to determine the transmission rate \( R(t) \) as a function of time \( t \) to optimize the following problem:

\[
\min_{R(t)} \int_0^\infty P(R(t)) \, dt \\
\text{s.t.} \int_{t_i}^{t_i'} R(t) \, dt = L, \quad i = 1, 2, \ldots, U \\
\int_{t_i}^{t_i'} R(t) \, dt \geq L, \quad i = 1, 2, \ldots, U \\
R(t) \geq 0,
\]  

(4)

where \( P(\cdot) \) is the power consumption given in equation (2). Considering that \( P(0) = 0 \) in our energy consumption model, this optimization problem equals the following one:

\[
\min_{R(t)} \int_0^\infty P(R(t)) \, dt \\
\text{s.t.} \int_{t_i}^{t_i'} R(t) \, dt = L, \quad i = 1, 2, \ldots, U \\
\int_{t_i}^{t_i'} R(t) \, dt \geq L, \quad i = 1, 2, \ldots, U \\
R(t) \geq 0,
\]  

(5)

in which the departure times are set to be the deadlines and the received data should be equal or larger than \( L \) bits for each user.

Here we divide the rate scheduling problem into an offline version and an online one, based on the knowledge of users’ arrivals (i.e. non-causal or causal user arrival information). The two scenarios are respectively analyzed in the following two sections.

\section*{III. Offline scheduling}

In the offline rate scheduling, the user arrivals are known in advance. The inter-arrival time between user \( i \) and user \( i + 1 \) is denoted by \( d_i \), where \( d_i = t_{i+1}' - t_i' \). Here we need to interpret the multicast session: when \( d_i \) is less than \( T \), which is the maximal delay of each user, both users at the two ends of \( d_i \) belong to the same session since they may have overlapped data to receive. We use this definition to maximize the number of users in a session under the constraints of user deadlines. For example, in Fig. 1 the arrows represent users’ arrivals and departures. User 1 belongs to session 1, while user 2, 3 and 4 are grouped into session 2. With this session definition, the transmission rate can be scheduled within each
session independently and thus the computational complexity is reduced. In each session, the users’ arrivals and departures will divide the session into several epochs, in each of which the number of users is constant. For instance, the second session includes epoch 2, 3, 4, 5 and 6 as illustrated in Fig. 1.

Proposition 1: The transmission rate should be constant in each epoch for optimality.

Proof: Let $t_1$ and $t_2$ be the starting point and terminal point of an epoch, respectively. The data transmitted in the epoch is assumed $L$ bits. The transmission rate is denoted by $R(t)$ as a function of time $t$ ($t_1 \leq t \leq t_2$). We assume the data transmitted from $t_1$ to $t_2$ equals $\int_{t_1}^{t_2} R(t)dt = L$ bits. The energy consumption is computed as follows:

$$E(R(t)) = \int_{t_1}^{t_2} P(R(t))dt.$$  \hspace{1cm} (6)

Utilizing the strict convexity of $P(\cdot)$ in equation (2), we obtain the following inequality:

$$E(R(t)) \geq (t_2 - t_1)P\left(\frac{L}{t_2 - t_1}\right),$$  \hspace{1cm} (7)

with equality if and only if $R(t) = \frac{\int_{t_1}^{t} R(t)dt}{t_2 - t_1} = \frac{L}{t_2 - t_1}$. So the optimal transmission rate is $R^*(t) = \frac{L}{t_2 - t_1}$ and is constant in the epoch.

Since the user arrivals follow a Poisson process, the probability that one user’s arrival coincide with another user’s departure is infinitely small. Therefore, when there are $N$ user arrivals in a multicast session, the number of epochs in this session will be $2N - 1$. We denote the duration of $j$-th epoch by $\epsilon_j$, $j = 1, 2, 3, \ldots, 2N - 1$ and the transmission rate in that epoch by $r_j$, $j = 1, 2, 3, \ldots, 2N - 1$. Therefore, the offline scheduling problem is:

$$\min_{r_1, r_2, \ldots, r_{2N-1}} \sum_{j=1}^{2N-1} P(r_j)e_j$$

In Fig. 2, the energy consumption of optimal algorithm is normalized by the energy consumption of the constant rate multicast. The dashed line is the result of Poisson process and the solid line shows the result of MMPP2. As the figure illustrates, the energy consumption is minimized when the
arrival rate is about 0.012 s\(^{-1}\) in both traffic models and about 13% of the energy is saved compared to the constant rate multicast. When the traffic is lower than this optimal point, the probability of data overlapping among users is small, so the energy saving gain is small. When the traffic is higher, more users overlap with others, which causes the session lasts for a longer period. As a result, on the one hand, it provides more chances for multicast; on the other hand, it makes the rate scheduling less efficient. In general, more than 10% of energy consumption is reduced under moderate and high traffic load.

IV. ONLINE SCHEDULING

For the online scheduling, the user arrivals are unknown in advance, and hence we need to determine the transmission rate based on current state and past information. However, the problem is hard to solve due to the multitude of variables and dimensions. Thus, we deal with a simplified case with only 2 users which can be formulated into a Markov decision process (MDP) and therefore can be solved by dynamic programming (DP). Based on the DP results, our heuristic lazy rate scheduling schemes are proposed.

A. MDP formulation and DP results

The online problem has the Markov property as we assume the memoryless traffic model. The effects of taking an action depend only on the current state and have no relationship with history, thereby making DP an effective way to solve the problem.

Firstly, consider a time-slotted system that include only two users: the first one triggers the multicast session at time 0 and the other will randomly join in the session at any timeslot \( t \) before the transmission is over. Both the users need to receive the \( L \) bits data before the deadline \( T \) and \( t + T \), respectively. We restrict the transmission rate to be chosen from integers. Our target is to minimize the energy consumption by scheduling the transmission rate \( r_k \) of each timeslot. Let \( L^b \) denote the data transmitted to user 1 before \( t \), i.e.,

\[
L^b = \sum_{k=1}^{t} r_k.
\]

After the arrival of user 2, there is no randomness in this system and the transmitter has to transmit data of length \( L^b \) to user 1 and data of length \( L \) to user 2; the data to user 2 includes the data to user 1. If we only consider user 2, the transmission rate should be \( \frac{L}{T} \) to minimize energy consumption. However, this may not meet the deadline of user 1. Thus the optimal transmission rate after the arrival of user 2 depends on the value of \( t \) and \( L^b \). If \( L - L^b \leq \frac{L}{T} (T - t) \), which means the transmission rate \( \frac{L}{T} \) can meet the deadline of user 1, the optimal transmission rate is \( \frac{L}{T} \) from timeslot \( t + 1 \) to \( t + T \). Otherwise, the transmission rate is \( \frac{L - L^b}{T - t} \) from timeslot \( t + 1 \) to \( T \), and is \( \frac{L}{T} \) from timeslot \( T + 1 \) to \( t + T \) for optimality. Therefore, given \( t \) and \( L^b \), the energy consumption after the arrival of user 2 is given by:

\[
C^l(t', L^b) = \begin{cases} 
T(2^{\frac{L}{T}} - 1), & \frac{L - L^b}{T - t} \leq \frac{L}{T} \\
(T - t')(2^{\frac{L}{T}} - 1) + t'(2^{\frac{L}{T}} - 1), & \text{otherwise}.
\end{cases}
\]

And the total energy consumption is represented as:

\[
\min_{r_1, r_2, \ldots, r_T} E \left\{ \sum_{k=1}^{t'} \left[ 2^{\frac{r_k}{T}} - 1 \right] + C^l(t', L^b) \right\} \quad \text{s.t.} \quad r_k = 0, 1, 2, 3, \ldots, \quad \text{for} \quad k = 1, \ldots, t'.
\]

The first part \( \sum_{k=1}^{t'} \left[ 2^{\frac{r_k}{T}} - 1 \right] \) in the objective function (10) accounts for the energy consumption before the arrival of user 2.

Let \( J_t(L^b) \) denote the minimum expected energy consumption from the \( k \)-th timeslot to the last timeslot with data of length \( L^b \) transmitted before the \( k \)-th timeslot, which is referred to as the cost-to-go function [14]. In the last timeslot, we have

\[
J_T(L^b) = \begin{cases} 
0, & L^b = L \\
+\infty, & L^b \neq L.
\end{cases}
\]

As a result, \( J_t(L^b) \) can be interpreted as:

\[
J_t(L^b) = E \left\{ \min(C(r_t) + J_{t+1}(L^b + r_t))(1 - L_t) + C^l(k, L^b)I_t \right\},
\]

where \( I_t \) is a 0-1 valued random variable representing whether user 2 joins in or not at timeslot \( k \). Specifically, \( I_t = 1 \) means user 2 joins in.

The timeslot when the user 2 joins is assumed to be geometric distributed with parameter \( Pr \). It denotes the probability with which \( I_t \) equals 1 in each timeslot. We set \( L \) to 1000 bits and the length of a timeslot to 1 second. Other parameters are given in Table I. Then, the optimal rate scheduling results under three levels of traffic are computed and shown in Fig. 3. The solid line shows the result when \( Pr = 0.1 \), representing a high user arrival probability; the dash-dotted line gives results when \( Pr = 0.01 \), representing a moderate user arrival probability; the dotted line corresponds to \( Pr = 0.001 \), representing a low user arrival probability. It shows that in the low traffic case, the transmission rate is constant, while in the high traffic case the rate increases exponentially. The comparison of three lines reveals that the more data should be transmitted in last timeslots before the deadline when the traffic load is high, and this motivates the design of our lazy rate scheduling.

B. Lazy scheduling schemes

The scenario in above subsection is far from practical. However, it implies that the transmission rate should be low at first and grow gradually before each deadline, such that more data is multicasted to potential users in the future. Based on this intuition of “Lazy Transmission”, 4 heuristic schemes are
proposed to determine ways in which the transmission rate grows. The performances are compared under the two traffic models given in section II.

To begin with, we introduce the first 3 heuristic schemes. As we mentioned before, the users join in system at time $t_1, t_2, \ldots$ and should finish reception before $t_1 + T, t_2 + T, \ldots$ As long as the multicast session is not over, there is an urgent user whose deadline is the earliest. So we consider this user’s remaining time and unreceived data with priority. As long as the multicast session is not over, there is an urgent user who finishes reception just by its deadline; the other users finish reception before their deadlines. Note the user who joins in the system after time $T_W$ has to setup a new multicast session now. Under different traffic load, the optimal $T_W$ can be found to achieve the best performance. Fig. 4 presents the four rate scheduling schemes when $L_{i+1} = 50$ Mbits, $r_{i+1} = 50$ s, $\beta = 0.2$, $T = 50$ s and $T_W = 30$ s.

At last, we show the performances of our lazy rate scheduling schemes in Fig. 5, considering Poisson process and MMPP$_2$ respectively. The parameter settings are given in Table I except that we discretize the time domain into timeslots as in the case of MDP. Since one timeslot corresponds to $1$ s, the impact caused by discretization is relatively small.

With different average arrival rate, the optimal parameter $\beta^*$ in scheme 1, 2, 3 and $T_W$ in scheme 4 are found by global search. All the values are normalized by the energy consumption of transmitting data with constant rate of $4$.

The four dashed lines in Fig. 5 show the performances of our schemes with Poisson user arrival model. The solid lines are the results of the MMPP$_2$ user arrival model. Comparing the two kinds of traffic models, we find that for MMPP$_2$ arrival process, our schemes have better performances under
Fig. 5. The performances of four heuristic schemes under two kinds of traffic models.

low traffic load and worse performances under high traffic load. The curves of the square-root, linear and quadratic lazy schemes start to reach their lowest values when the equivalent arrival rate $\lambda$ is around 0.014 s$^{-1}$. At this point, the high-level lazy scheme saves energy by 8% for both traffic models. Beyond this arrival rate, more users coexist in a multicast session, which causes the session to last much longer. Thus the cost of increasing the transmission rate offsets the revenue brought by lazy rate scheduling. The 0-1 lazy rate scheduling scheme can save energy by 10% when the arrival probability is near 0.02 s$^{-1}$ and this outperforms the other three heuristic schemes greatly. However, when user arrival rate is low, for example 0.003 s$^{-1}$, the 0-1 lazy scheme will consume more energy compared even with the constant rate scheme. When the arrival rate is higher than 0.02 s$^{-1}$, the 0-1 lazy scheme can further reduce energy consumption. Another advantage of the 0-1 lazy scheme is that the transmission rate does not vary from timeslot to timeslot. Therefore, when the multicast network is under high traffic load, it is a preferable option to adopt 0-1 lazy rate scheduling scheme considering both energy-efficiency and implementation complexity.

V. Conclusions

In this paper, we aim at reducing energy consumption for QoS guaranteed download services in wireless multicast networks via rate scheduling. In traditional multicast system, the transmission rate is constant. However, we show that it is far from optimal if we consider energy-efficiency. Therefore, rate adaptive scheme is designed to schedule the transmission rate according to the current state of the system. Based on the non-causal and causal information about user arrivals, the problem is formulated into offline and online scheduling versions. The optimal offline scheduling algorithm can save energy by nearly 15%. For the online scheduling, the problem is simplified to an MDP with 2 users. Based on the solutions to this simplified problem, we propose heuristic schemes, i.e., lazy rate scheduling. The performance evaluation shows that our lazy rate scheduling can save up to 10% energy.

ACKNOWLEDGMENT

This work was supported in part by the National Basic Research Program of China (973 Program: No. 2012CB316001), the Nature Science Foundation of China (No. 61021001, No.60925002) and Hitachi R&D Headquarter.

REFERENCES