Efficient Group-Based Multimedia-on-Demand Service Delivery in Wireless Networks

Luan Huang, Kar Ann Chew, Member, IEEE, Shyamalie Thilakawardana, Yajian Liu, Student Member, IEEE, Klaus Moessner, Member, IEEE, and Rahim Tafazolli, Member, IEEE

Abstract—Recently, an upsurge of interest has been observed in providing multimedia on-demand (MoD) services to mobile users over wireless networks. Nevertheless, due to the rapidly varying nature of mobile networks and the scarcity of radio resources, the commercial implementation is still limited. This paper presents an efficient group-based multimedia-on-demand (GMoD) service model over multicast-enabled wireless infrastructures, where users requesting the same content are grouped and served simultaneously with a single multicast stream. The grouping is fulfilled through a process named “batching”. An analytical model is derived to analyse a timeout-based batching scheme with respect to the tradeoff between user blocking probability and reneging probability. Based on the deduced analytical model, an optimal timeout-based batching scheme is proposed to dynamically identify the optimal tradeoff point that maximizes the system satisfaction ratio given a particular system status. The proposed scheme is evaluated by means of simulation and compared with two basic batching schemes (timeout-based, size-based), and two hybrid ones (combined-for-profit, combined-for-loss). The simulation results demonstrate the proposed approach can ensure significant gains in terms of user satisfaction ratio, with low reneging and blocking probabilities.

Index Terms—Batching, DVB, MBMS, MoD, resource management.

I. INTRODUCTION

The cellular network has been developed and evolved well for offering on-demand services to mobile users, such as voice, short message and Internet access. Nevertheless, users’ growing demands for bandwidth-intensive multimedia services make it more attractive to offer multimedia-on-demand (MoD) services to mobile users, such as video-on-demand and music-on-demand. In the on-demand service mode, a user will place a request for a service directly through the network. The service is then delivered to this user according to the user preference as well as the resource availability and service requirements.

MoD services are traditionally delivered via unicast transmissions, by which, data are sent separately to individual recipients. Hence, a traffic increase implies additional consumption of the network resources. In recent years, there has been great interest in utilizing multicast technology in both wired and wireless networks due to its cost-effective transmission [1]–[10]. Unlike unicast, in multicast, only one copy of data is transmitted from the source to the multiple recipients in the multicast group. Therefore, this leads to less resource consumption in comparison to unicast.

Recent standardization efforts in wireless networks, such as MBMS [11], aim to support multicast in 3G mobile networks and the standardization effort of Digital Video Broadcasting transmission system for Handheld terminal (DVB-H) gives another possibility for utilizing multicast in broadcasting networks to handheld devices [12]. In addition, effective cooperation between cellular and broadcasting networks enables the cost-effective provisioning of on-demand services [13]. In this cooperating system, the 3G networks is used to offer a return channel to collect user requests for services, and the DVB network offers efficient service delivery to users by using multicasting or broadcasting delivery methods. In our previous work, we proposed an interworking architecture encompassing MBMS and DVB networks [14]. Relevant functional entities have also been defined, together with the interfaces, signaling flows and information primitives. Although, a lot of conceptual work has recently for making the existing GMoD systems more efficient...
and more timely, such as the batching method [1]–[4], the piggybacking method [5], and the patching method [6]. These techniques are complementary, and hence can be combined into hybrid strategy to further improve efficiency [7], [8]. Out of these methods, the batching method is more computationally efficient, and avoids additional resource requirement, such as buffer capacity in the mobile terminals. Therefore, the study in this paper focuses on the batching method and attempts to optimize this method in a dynamic wireless environment.

The two basic batching methods in wireline networks are known as timeout-based batching and size-based batching [1], depending on whether the users are batched for a fixed period of time, or are batched until a sufficient number of requests are collected. These two basic batching schemes constitute the foundation for the later extended schemes. For instance, two hybrid methods have been proposed by Chan [2], namely, combined-for-profit batching and combined-for-loss batching. Furthermore, the timeout-based batching scheme has been analysed from a business perspective in [3]. In both [2] and [3], analytical models were derived for performance evaluation of the timeout-based batching scheme. However, they consider infinite network resource and study a single content [2], or, the user’s reneging behavior is not considered [3]. Obviously, it is suitable neither for the resource-limited wireless environment, nor for practical implementation. In [4], a new batching method, called delay-aware broadcasting is proposed, making use of the exact delay tolerance of each user. This is however difficult to implement due to the difficulty of knowing the user’s true delay preference.

This paper enhances the traditional analysis models of the basic timeout-based batching scheme by considering the blocking probability caused by the limited wireless resource, and furthermore we consider multiple contents and the behavior of user reneging. Based on these studies, an analytical model is derived to study the performance of the timeout-based batching scheme. Moreover, we propose a novel optimal timeout-based batching scheme to maximize the percentage of served users with respect to the total users, adaptive to the available resources and traffic loads in the system. Streaming real-time application is considered in this work as it is envisioned as an ideal driver for the next generation mobile systems [16].

This paper is structured as follows. First, the system model and four traditional batching schemes are introduced in Section II. In Section III, a theoretical model of the basic timeout-based scheme is presented by considering the blocking probability, reneging probability, user satisfaction ratio and multiple contents. Based on this model, our optimal timeout-based scheme is proposed. The simulation results are presented and discussed in Section IV. Finally, Section V concludes this paper.

II. REVIEW ON BATCHING SCHEMES

A. System Model

The system model of the GMoD provisioning is illustrated in Fig. 1.

Fig. 1. System model of the GMoD provisioning.
this content. Furthermore, after batching, the system sends a resource request for serving the content on behalf of all the remaining user requests in the batching queue. This resource request competes with other resource requests for a limited multicast resource pool containing all the multicast channels in a wireless cell. The arrival rate of the resource request for content \(i\) is denoted as \(\lambda_i\). If there is a free multicast channel available, the batch of users that sends the resource request is served simultaneously. Otherwise, the batch of users is lost from the system without being served. The system model will be further explained with a mathematical model in Section III.

### B. Traditional Schemes

Four batching schemes were proposed in [1], [2]. They are two basic schemes (timeout-based, size-based) and two hybrid schemes (combined-for-profit, combined-for-loss). In these schemes, the requests for the same content are batched for a specific batching duration, and then are served with a single multicast stream. This batching duration can be identified by the period of user’s waiting time in the batch (timeout-based scheme), the number of users’ requests in the batch (size-based scheme), or by both waiting time and number of requests (hybrid schemes). In either of these schemes, the batching duration can cause users reneging from their request or the batch due to the delay in content delivery. These four batching schemes are described below briefly.

**Timeout-based**: batching duration is identified by the period of the first user’s waiting time \(W\) referred to as the batching time in order to guarantee the maximum service delay. Requests are batched until the batching time is reached. The timer is set when a user agrees to wait \(W\) minutes and the subsequent users for the same content join the waiting line.

**Size-based**: batching duration is identified by the number of users \(N\) referred to as the batching size in order to guarantee the number of users served as a batch. Requests are batched until the batching size threshold is reached. Hence, the user’s waiting time depends on the rate of request arrivals.

**Combined-for-profit / loss**: batching duration is identified by both of \(W\) and \(N\). In the combined-for-profit scheme, requests are batched until both \(N\) users are collected and the first user in the batch waits no less than \(W\) time. This scheme aims to maximize the “profit” (e.g., the number of served users). In the combined-for-loss scheme, requests are batched until either \(N\) number of users is collected or the first user in the batch has waited \(W\) time units. This scheme aims to minimize the service delay, thus the reneging probability.

### III. ANALYSIS AND OPTIMIZATION

Three metrics of the performance in the basic timeout-based scheme will be formulated. They are the reneging probability \(P_{ren}\), blocking probability \(P_{bk}\) and satisfaction ratio \(P_s\), where \(P_{ren}\), \(P_{bk}\) and \(P_s\) are, respectively, the ratio of the number of reneged users due to the delayed service set-up, blocked users due to the insufficient resources and served users, with respect to the total number of users in the system. Obviously, \(P_s = 1 - P_{bk} - P_{ren}\). Important notations used in this paper are summarized in Table I.

#### A. Analytical Model

Users are able to request \(M\) independent contents according to a certain probability. Let \(a_i\) be the probability that a request chooses content \(i\), and let \(a_1 \geq a_2 \geq \cdots \geq a_M\) without loss of generality.

The arrivals of users’ requests are assumed to be Poisson process with a rate of \(\lambda\). And each arrival of Poisson \((\lambda)\) can select one of the \(M\) contents. Therefore, the Poisson \((\lambda)\) can be decomposed into \(M\) independent Poisson processes: Poisson \((\lambda a_1)\), Poisson \((\lambda a_2)\) and Poisson \((\lambda a_M)\), each modeling the arrival of requests for content \(i\).

Each user is willing to wait for a period of time \(U \geq 0\) to view the desired content. If this content is not displayed by then the request reneges. \(U\) is a random variable with its cumulative distribution denoted by a user reneging function \(R(u) = P(U < u)\) with a mean \(\tau\).

Let \(\lambda_i\) be the throughput of content \(i\), in terms of the number of remaining requests in unit time, which do not renage during batching. \(\lambda_i = \frac{\overline{N}_i}{\overline{T}_i}\), where \(\overline{N}_i\) is the average batch size in number of requests of content \(i\) and \(\overline{T}_i\) is the average period between two consecutive copy of content \(i\) delivery time. In [2], it is derived that

\[
\overline{T}_i = \frac{1}{(1 - R(W))\lambda a_i + W} \quad (1)
\]

\[
\overline{N}_i = 1 + \lambda a_i W \int_0^W (1 - R(x)) \frac{dx}{W} \quad (2)
\]

We aggregate \(M\) arrival processes into one process. The total throughput \(\lambda'\) is then given by

\[
\lambda' = \sum_{i=1}^M \lambda_i = \sum_{i=1}^M \frac{\overline{N}_i}{\overline{T}_i} \quad (3)
\]

So, the reneging probability \(P_{ren}\) is given by

\[
P_{ren} = 1 - \lambda'/\lambda \quad (4)
\]
Based on the (1)–(4), we can derive that $P_{\text{ren}}$ has an upper limit with $\lambda$, i.e.,

$$\lim_{\lambda \to +\infty} P_{\text{ren}}(\lambda) = 1 - \frac{\int_{0}^{W} (1 - R(x)) \, dx}{W}$$

(5)

After batching, remaining requests compete for a total of $C$ multicast radio channels for delivering multicast contents. Requests are blocked if there is no channel available.

Given the total $C$ multicast channels, the GMoD system can be modeled as an $M|G_1/C/C$ queue. Let $P_{\text{blk}}^d$ be the probability that the total channels are occupied, the blocking probability $P_{\text{blk}} = P_{\text{blk}}^d \cdot (1 - P_{\text{ren}})$. $P_{\text{blk}}^d$ can be calculated using Erlang B Formula [17] and is given as following.

$$P_{\text{blk}}^d = \frac{\lambda_s(C\lambda_s d)^C}{C! \sum_{i=0}^{C} \frac{(C\lambda_s d)^i}{i!}}$$

(6)

where $\lambda_s$ is the arrival rate of total resource requests for serving desired contents. Let $\lambda_s^i$ be the request arrival rate for content $i$, $\lambda_s$ is given by

$$\lambda_s = \sum_{i=1}^{M} \lambda_s^i = \sum_{i=1}^{M} \frac{1}{i}$$

(7)

Based on the equation, we can derive that $P_{\text{blk}}^d$ has an upper limit with $\lambda$ i.e.,

$$\lim_{\lambda \to +\infty} P_{\text{blk}}^d(\lambda) = \frac{\lambda_s^C (Md/W)^C}{C! \sum_{i=0}^{C} \frac{(Md/W)^i}{i!}}$$

(8)

Since $P_{\text{ren}}$ also has an upper limit, the upper limit of $P_{\text{blk}}^d$ and thereby $P_s$ exist, i.e.,

$$\lim_{\lambda \to +\infty} P_s = 1 - \lim_{\lambda \to +\infty} P_{\text{blk}}^d \cdot (1 - \lim_{\lambda \to +\infty} P_{\text{ren}}) = \lim_{\lambda \to +\infty} P_{\text{ren}}$$

(9)

These results indicate that $P_{\text{ren}}$, $P_{\text{blk}}^d$, and $P_s$ will not be dependent on $\lambda$ when $\lambda$ increases to a certain value.

To study the performance of the reneging probability, the exponential user reneging function in [1], [2] is adopted in this paper, i.e.,

$$R(u) = \begin{cases} 0, & \text{if } 0 \leq u \leq U_{\min} \\ 1 - e^{-(u-U_{\min})/\tau}, & \text{otherwise}. \end{cases}$$

(10)

where $U_{\min}$ is the minimum time for which users are always willing to wait.

### Table II

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>1000 requests/hour</td>
</tr>
<tr>
<td>$\tau$</td>
<td>10 minutes</td>
</tr>
<tr>
<td>$d$</td>
<td>15 minutes</td>
</tr>
<tr>
<td>$l_{\min}$</td>
<td>1 minute</td>
</tr>
<tr>
<td>$C$</td>
<td>20</td>
</tr>
<tr>
<td>$M$</td>
<td>(e.g., news, travel info, sports, music, music info)</td>
</tr>
<tr>
<td>$\theta$</td>
<td>0.271</td>
</tr>
<tr>
<td>$W$</td>
<td>5 minutes</td>
</tr>
</tbody>
</table>

Zipf distribution is adopted to model the probability of choosing a particular content [20]. The probability of choosing the $k$th content is given by:

$$p_k = \frac{1}{k^{(1-\theta)} \sum_{i=1}^{M} k^{(1-\theta)}}$$

(11)

where, $\theta$ is the parameter that determines the shape of the distribution. A small $\theta$ corresponds to a more severe discrimination of requests among contents and indicates that some contents are desired more frequently than the others. As in [3], $\theta = 0.271$ is assumed in this work.

Table II shows the default parameter values used in the analysis and simulation unless otherwise stated.

### B. Simulation Model

A simulator was constructed using OPNET [19] to justify the derived analytical model and evaluate the proposed optimal timeout-based batching scheme by comparing with four conventional batching schemes, namely, two traditional basic schemes (timeout-based, size-based) and two traditional hybrid schemes (combined-for-profit, combined-for-loss). The assumption in the analytical model is adopted by the simulation model unless otherwise stated. This simulator consists of a wireless cell (e.g., DVB cell) with limited available multicast radio channels. Users are uniformly distributed in the cell. Each user randomly sends a request for a particular content at any one time. The duration of the content is determined based on an Exponential distribution. We consider a mixed traffic scenario where the GMoD service and the other services (e.g., TV broadcasting and datacasting) share and compete for a common and limited resource pool. As a result, the network resource available for the GMoD service fluctuates over time, as denoted by C multicast channels.

### C. Analytical Results

Both the results obtained from computer simulations and analytical model are shown in Fig. 2. We observe that the performance depends highly on the selection of batching time $W$. Thus, finding an appropriate $W$ is a crucial issue. $P_s$ first increases to reach a maximum, and then decreases rather slowly as $W$ increases. Maximum $P_s$ is achievable given certain system status, which includes service quality constraints (e.g., $P_{\text{ren}}$...
and $P_{dk}$ constraint), traffic information (e.g., content duration, request arrival rate), user profile (e.g., reneging function), and resource availability. The $W$ at which the maximum $P_s$ is achieved is denoted as the optimal batching time $W_{op}$. We can see a trade-off between $P_{ren}$ and $P_{dk}$. $P_{ren}$ increases rather linearly when $W$ increases as more users renege with the longer batching time, whereas $P_{dk}$ decreases due to the more efficient resource utilization. Also we can see that $P_{ren}$ varies more gradually than $P_{dk}$ until $P_{dk}$ decreases to reach 0 as $W$ increases. This indicates that $P_{dk}$ is a dominant factor affecting $P_s$ when resource is not enough; otherwise, $P_{ren}$ becomes dominant. The lower bound of $W_{op}$ is determined by the constraint of the maximum $P_{dk}$ while the upper bound of $W_{op}$ is determined by the constraint of the maximum $P_{ren}$. For instance, given the constraints $P_{ren} \leq 0.1$ and $P_{dk} \leq 0.05$, we can use $W_{op} = 4$ minutes.

Fig. 3 shows that the available resource $C$ has a significant impact on the system performance. Fig. 3(a) shows that as $C$ increases, $P_{dk}$ reduces as more resources are available. However, $P_{ren}$ is completely independent of $C$. This is consistent with the analytical model. Furthermore, it can be seen from Fig. 3(b) and Fig. 3(c) that $W_{op}$ can be achieved at low value of $W$, especially when $C$ is high. When $C$ increases till $P_{dk} = 0$, $C$ has no impact on the performance. Hence, $W_{op}$ can be kept as low as possible in order to minimize $P_{ren}$. As $C$ decreases, more users have to be batched in a multicast channel in order to efficiently utilize more scarce resources, and hence $W_{op}$ increases. This indicates that when resource is sufficient, a short $W$ is required in order to minimize $P_{ren}$; otherwise, a long $W$ is required in order to maximize resource utilization efficiency.

Fig. 4 shows that $\lambda$ has not much influence on the performance, especially when $\lambda$ is high. This result justifies the analysis in the analytical model. Fig. 4(a) shows that as $\lambda$ increases, $P_{dk}$ and $P_{ren}$ first decreases and increases respectively, and then both change rather slowly. Therefore, a constant $W_{op}$ can be used for a wide range of $\lambda$ especially when $\lambda$ is high, which can be seen in Fig. 4(b) and Fig. 4(c). For instance, to maximize $P_s$ while maintaining the maximum $P_{ren}$ at around 0.2 and $P_{dk}$ at around 0.05, we can use $W = 5$ minutes for $\lambda$ ranging from 1000 to 10000 requests/hour. And the minimum $\lambda$ at which the $P_s$ keeps stable can be calculated by finding the minimum $\lambda$
from the solutions of the function \( \lim_{\lambda \to \infty} P_a - P_a(\lambda) = \varepsilon \) \((\varepsilon = 0.01)\). As shown in Fig. 4(a), as \( \lambda \) gets smaller, \( P_{\text{ren}} \) becomes a dominant factor affecting \( P_a \), so a shorter \( W \) is required to reduce \( P_{\text{ren}} \). For instances, Fig. 4(b) and Fig. 4(c) indicate that when \( \lambda \) varies from 1000 to 100 requests/hour, we can reduce \( W_{\text{op}} \) from 5 to 2 minutes. This result indicates that in the case of a high \( \lambda \), it is not necessary to adapt \( W \) to \( \lambda \). However, when \( \lambda \) is low, a shorter \( W_{\text{op}} \) is required in order to minimize \( P_{\text{ren}} \).

We continue studying the performance of the basic timeout-based scheme when varying the user’s mean waiting time from 1 to 10 minutes. Fig. 5 presents the performance of \( P_a \), \( P_{\text{blk}} \), and \( P_{\text{ren}} \).

It can be seen from Fig. 5(a) that the mean waiting time \( \tau \) has different impact on \( P_{\text{blk}} \) and \( P_{\text{ren}} \). When \( \tau \) increases, users are more likely to stay in the batching queue without reneging. Hence, smaller \( P_{\text{ren}} \) is achieved. However, as defined in the analytical model, \( P_{\text{blk}} \) is the percentage of blocked users excluding reneged users, with respect to the total users in the system. Therefore, \( P_{\text{blk}} \) is negative proportional to the \( P_{\text{ren}} \). Moreover, as indicated in (1), (6), (7), with larger mean waiting time, users are less likely to renge, which results in more resource requests competing for limited resources, hence increasing \( P_{\text{ren}} \). Therefore, \( P_{\text{blk}} \) varies in a different way from \( P_{\text{ren}} \), and is positive proportional to \( \tau \). In addition, it can be seen that \( P_{\text{blk}} \) varies more gradually than \( P_{\text{ren}} \). Hence, \( P_a \) increases as \( \tau \) increases. This is also demonstrated in Fig. 5(b). Moreover, it can be observed that \( \tau \) has reduced impact on the performance of \( P_a \),
Fig. 6. Impact of the shape parameter $\theta$ of Zipf distribution on system performance. (a) Satisfaction ratio, blocking probability and reneging probability vs. the shape parameter $\theta$ of Zipf distribution, with $W = 5$ minutes; (b) performance difference in terms of satisfaction ratio, blocking probability and reneging probability, with respect to $\theta = 0$; (c) satisfaction ratio vs. batching time.

Table III

Mapping Table in Optimal Timeout-Based Scheme, with $\tau = 10$ minutes. (A) Mapping Table of $C$ to $W_{opt}$ with $\lambda = 1000$ request/hour: (b) Mapping Table of $\lambda$ to $W_{opt}$ with $C = 20$

<table>
<thead>
<tr>
<th>Number of available channels $C$</th>
<th>Optimal batching time $W_{opt}$ (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>35</td>
<td>2</td>
</tr>
<tr>
<td>40</td>
<td>2</td>
</tr>
</tbody>
</table>

(B)

<table>
<thead>
<tr>
<th>Request arrival rate $\lambda$ (request/hour)</th>
<th>Optimal batching time $W_{opt}$ (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>400-1900</td>
<td>4</td>
</tr>
</tbody>
</table>

$\tau$ lead to the same $P_s$. This is because every user has a minimum waiting time threshold $U_{min} = 1$ minute, for which users will stay in the queue without reneging. Furthermore, as $\tau$ gets smaller, shorter period of the batching time is required to achieve the maximum $P_s$. It is because as $\tau$ gets smaller, $P_{ren}$ increases dramatically, while $P_{blk}$ reduces gradually, that shorter batching time is needed to reduce $P_{ren}$.

The performance when varying the shape parameter $\theta$ of the Zipf distribution is also studied. Fig. 6 presents the performance of $P_s$, $P_{blk}$, and $P_{ren}$.

It can be seen from Fig. 6(a) and Fig. 6(c) that $\theta$ has no obvious influence on the performance of $P_s$, $P_{blk}$, and $P_{ren}$. However, if the performance is enlarged by measuring the performance difference with respect to $\theta = 0$, the impact of $\theta$ can be observed. As shown in the Fig. 6(b), the system generates larger $P_{blk}$ and $P_{ren}$, thereby smaller $P_s$, as $\theta$ increases. This is because the smaller value of $\theta$ means the users’ requests concentrate on fewer contents, which would make the batching more effective. These results above are consistent with those depicted in [3], where greater impact of $\theta$ can be observed on the performance of system revenue by associating a price to a certain content, or the measurement of the blocking probability of mixed traffic comprising batching services and traditional interactive services.

D. Optimal Timeout-Based Scheme

Based on our analytical model, $W_{opt}$ can be calculated given a certain system status including the service quality, traffic information, user reneging function and resource availability. Therefore, we can come up with a mapping table from a given system status to $W_{opt}$, such as in Table III and Table IV. With the mapping table, the network or service providers can keep track of the system current status, in particular the available resources, to determine $W_{opt}$ at that moment. Users’ requests for contents are then scheduled and served based on the $W_{opt}$. Considering that $W_{opt}$ does not depend much on the request arrival rate $\lambda$, when $\lambda$ is high, the size of the mapping table can be decreased by mapping a range of $\lambda$ to a particular $W_{opt}$.
In a resource-limited wireless communications system, $P_{dlk}$ is the key concern when evaluating the grade of service (GoS) of on-demand services. This is because it is more annoying for the users to wait but eventually not being served than to decide not to wait and thereby reneging from the request. Moreover, effective methods to increase the users’ delay tolerance, such as decreasing the service price, are easy to implement [18], hence the $P_{ren}$ can be effectively controlled. Therefore, we set $P_{dlk}$ as a priority constraint to $P_{ren}$. In our optimal batching scheme, the $W$ is optimized to maximize the satisfaction ratio $P_s$ subject to constraints of the maximum $P_{dlk} (5\%)$ and $P_{ren} (30\%)$.

### IV. EVALUATION OF OPTIMAL TIMEOUT-BASED SCHEME

#### A. Algorithm Parameters

We evaluate the proposed optimal timeout-based batching scheme by simulation. In the simulation, the optimal batching time $W_{op}$ adapts to the time varying resource availability and request arrival rate. Two experiments are performed with respect to different values of the user’s mean waiting time $\tau$ (i.e., 10 minutes and 3 minutes). The mapping tables with different $\tau$ are presented in Table III and Table IV respectively. Each mapping table contains the mapping of number of available channels and request arrival rate to the optimal batching time $W_{op}$. As a result, if the number of available channels is low, the batching time is set as long as possible to improve resource utilization by merging enough users in a group. Otherwise, the system will shorten the batching time in order to avoid too many users reneging from the batch. In contrast, given a certain number of available channels $C$, the batching time does not change much for a wide range of $\lambda$. In addition, comparing Table III and Table IV, it can be seen that a smaller $W_{op}$ is applied to a smaller $\tau$, under the same conditions. Furthermore, for a smaller $\tau$, more available channels are required so as to meet the constraint requirements of $P_{dlk}$ and $P_{ren}$ at the same time.

In the traditional batching schemes (i.e., the basic and hybrid schemes), the resource constraint is not considered. Their batching parameters, namely batching time $W$ and batching size $N$, are chosen from its respective constituent batching schemes (i.e., $W$ in the basic timeout-based scheme and $N$ in the basic size-based scheme) “optimized” independently for a target arrival rate and remain constant for other rates [2]. Therefore, we can characterize these traditional schemes as static resource management [21] in the sense that their batching parameters do not adapt to the variation of actual system status, such as the resource availability.

We evaluate the traditional schemes in a scenario where the cell capacity is limited, and moreover, the available resources are time-varying. The batching parameters of traditional schemes are configured based on the target user request arrival rate and the heaviest load of background traffic in the cell, as the methodology employed by the static resource allocation [22] in traditional cellular networks.

#### B. Simulation Results and Discussions

1) **Mean Waiting Time** $\tau = 10$ minutes: Fig. 7 presents $P_s$, $P_{dlk}$, and $P_{ren}$ as $C$ varies from 10 to 40 with $W = 9$ minutes, $N = 23$ in traditional schemes and $\lambda = 1000$ requests/hour. The parameters $W$ and $N$ are independently chosen in traditional schemes, given the target available channels $C = 10$.

Fig. 7(a) shows that the proposed optimal timeout-based scheme gives higher $P_s$ than other schemes for investigated values of $C$, especially when $C$ is large. This is because the optimal scheme chooses the most appropriate batching time for each $C$ so as to maximize $P_s$. In contrast, in traditional schemes, the batching parameters are dimensioned for the worst case (i.e., smallest $C$). These parameters result in a longer batching duration, which is not appropriate for non-worst case $C$. In addition, it can be seen that $P_s$ mainly depends on $P_{ren}$. This is because $P_{dlk}$ is so small that the $P_{ren}$ is dominant over the $P_{dlk}$ to determine the $P_s$.

Fig. 7(b) shows that the optimal scheme improves $P_{ren}$ as $C$ gets larger by choosing shorter batching time. However, the traditional schemes use the same batching parameters (i.e., batching time and batching size) as $C$ varies. Hence, the user’s waiting time and thereby $P_{ren}$ keeps constant in relation to $C$. In other words, it is not optimized and improved with respect to $C$. In addition, it can be seen that for all investigated values of $C$, the combined-for-loss scheme outperforms the other traditional schemes in reducing $P_{ren}$, especially, at the worst case $C$ (i.e., $C = 10$ channels), it outperforms the optimal scheme in terms of $P_{ren}$. This is because of the extra consideration of the batching size that is used together with the batching time to limit the batching duration in the combined-for-loss scheme. As a result, the batching duration is reduced compared to the other schemes.

Fig. 7(c) shows that although the two basic schemes and the combined-for-profit scheme give lower $P_{dlk}$ than the optimal scheme, the optimal scheme still keeps $P_{dlk}$ under 0.05 for all investigated $C$. In addition, it can be observed that the curve of the optimal scheme shows two peaks at $C = 20$ and $C = 35$. As seen in Table III, these peaks appear when the optimal batching time changes. However, it is not always the case, such as $C = 15$ and $C = 30$. This is due to the diverse impact of $C$ and $W$ on the blocking probability. As $C$ increases, the conventional

<table>
<thead>
<tr>
<th>Number of available channels $C$</th>
<th>Optimal batching time $W_{op}$ (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>3</td>
</tr>
<tr>
<td>30</td>
<td>2.5</td>
</tr>
<tr>
<td>35</td>
<td>2</td>
</tr>
<tr>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>45</td>
<td>1.5</td>
</tr>
<tr>
<td>50</td>
<td>1.5</td>
</tr>
<tr>
<td>55</td>
<td>1.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Request arrival rate $\lambda$ (request/hour)</th>
<th>Optimal batching time $W_{op}$ (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>400-700</td>
<td>1.5</td>
</tr>
<tr>
<td>1000-1900</td>
<td>2</td>
</tr>
</tbody>
</table>
schemes with fixed $W$ give smaller blocking probability. However, for the optimal timeout-based scheme, as $C$ increases, the $W$ decreases (see Table III). As a result, larger $C$ causes smaller blocking probability, whilst smaller $W$ leads to larger blocking probability. If the impact of $C$ is greater than that of $W$, $P_{\text{hk}}$ reduces, such as when $C$ changes from 10 to 15. Otherwise, $P_{\text{hk}}$ increases, such as when $C$ changes from 15 to 20, and from 30 to 35. When $C$ continues increasing from 20 to 25, or from 35 to 40, $P_{\text{hk}}$ gets smaller, because the same value of $W$ is applied at $C = 20$, $C = 25$, and at $C = 35$, $C = 40$ respectively.
Another thing we can see is that the combined-for-loss scheme gets the highest $P_{bk}$ at the targeted $C$ due to the shortest batching duration, which weakens its improvement in $P_{ren}$.

Fig. 8 presents performances of $P_s$, $P_{bk}$, and $P_{ren}$ when $\lambda$ changes from 400 to 1900 requests/hour with $W = 4$ minutes, $N = 14$ in traditional schemes and $C = 20$. The parameters $W$ and $N$ are independently chosen in traditional schemes, given the target arrival rate $\lambda = 1000$ requests/hour.

It can be seen that two timeout-based schemes including optimal scheme and basic scheme offer rather flat $P_s$, $P_{bk}$ and $P_{ren}$ as a function of $\lambda$, while in the other schemes these metrics can vary quite significantly with $\lambda$. Fig. 8(a) shows that the optimal scheme and the basic timeout-based scheme perform the same because they use the same duration of the batching time. Furthermore, these two timeout-based schemes give more consistent and higher $P_s$ than other schemes. As $\lambda$ goes higher than the target rate, the basic size-based scheme gives lower $P_s$. It is because in the size-based scheme, users are batched until $M$ users are collected. As $\lambda$ increases, shorter time is required to aggregate $M$ users, which increases the traffic load in the system, thus increasing $P_{bk}$, as shown in Fig. 8(b). But $P_{ren}$ is reduced due to the shorter batching time, as presented in Fig. 8(c). As the $P_{bk}$ is dominant to $P_{ren}$, $P_s$ goes down. As $\lambda$ drops, longer time is required to aggregate $M$ users in a batch, which leads to a smaller $P_{bk}$ but larger $P_{ren}$. As a result, $P_s$ is deduced due to the dominance of $P_{ren}$, which can be seen in Fig. 8(a).

The combined-for-profit scheme first operates according to the timeout-based scheme when $\lambda$ is high, and according to the size-based scheme when $\lambda$ drops. In contrast, the combined-for-loss scheme operates as the timeout-based scheme when $\lambda$ is low, and the size-based scheme when $\lambda$ increases. These results are consistent with those in [2].

2) Mean Waiting Time $\tau = 3$ minutes: Fig. 9 presents $P_s$, $P_{bk}$, and $P_{ren}$ as $C$ varies from 25 to 55 with $W = 3$ minutes, $N = 10$ in traditional schemes and $\lambda = 1000$ requests/hour. The parameters $W$ and $N$ are independently chosen in traditional schemes, given the target available channels $C = 25$.

It can be seen that in comparison to Fig. 7, the smaller value of $\tau$ does not influence a lot on the performance, as more channels are used. The optimal timeout-based scheme outperforms the other schemes in terms of $P_s$. Furthermore, among the conventional schemes, the combined-for-loss scheme achieves the largest $P_s$. Hence, the simulation results in Fig. 9 are consistent to those in Fig. 7.

Fig. 10 presents performance of $P_s$, $P_{bk}$, and $P_{ren}$ when $\lambda$ changes from 400 to 1900 requests/hour with $W = 2$ minutes, $N = 7$ in traditional schemes and $C = 40$. The parameters $W$ and $N$ are independently chosen in traditional schemes, given the target arrival rate $\lambda = 1000$ requests/hour.

It is observed from Fig. 10 that the trend of the performance curve of all the evaluated schemes is similar to that when $\tau = 10$ minutes. presented in Fig. 8. For instance, as shown in Fig. 10(a), two timeout-based schemes generally achieve rather flat and better $P_s$, in comparison to other schemes. However, in the case of small request arrival rate, i.e., 400 requests/hour and 700 requests/hour, the combined-for-loss scheme outperforms the basic timeout-based scheme, although it is outperformed by the optimal scheme. As shown in Fig. 10(c), it is because of the extra control parameter, batching size used in the combined-for-loss scheme, that it gets the smallest $P_{ren}$ among all
time when the request arrival rate is small. Hence, the optimal scheme outperforms all the conventional schemes in $P_s$.

V. CONCLUSIONS

This paper studied five batching schemes including four traditional schemes and one proposed novel approach to enable the resource efficient MoD service delivery in resource-limited wireless networks. We derived an analytical model for the performance evaluation of basic timeout-based scheme considering some practical wireless resource constraints. We found that there is a tradeoff between the blocking probability and the reneging probability. Also we found that there exists an optimal batching time that can strike this balance. Furthermore it was found that the change of request arrival rate has no much impact on the system performance, such as the satisfaction ratio, blocking and reneging probability, especially when the request arrival rate is high. On the other hand, the change of available resources gives different optimal batching time.

An optimal timeout-based scheme was proposed based on this analytical model to maximize the satisfaction ratio. In this approach, the batching duration time adapts to a certain system status characterized by service quality constraints, traffic information, user profile and resource availability. Simulation results showed improved performance using the proposed optimal scheme when the system status varies. In addition, we re-evaluated the traditional batching schemes, but this time with wireless resource limitation also being considered. Our results show that to maximize the satisfaction ratio, when the request arrival rate changes, the basic timeout-based scheme outperforms the other traditional schemes, while the combined-for-loss is better than the other traditional schemes as the available resources changes.

The proposed optimal scheme provides a superior solution to service providers or network operators to offer multicast-based MoD service over resource-limited wireless networks. Furthermore, under varying resource limitations, service providers and network operators can still get superior performance by adapting the batching time to the changes accordingly so as to make an optimal balance between blocking users and users reneging, in other words, between profitability and quality of service.

ACKNOWLEDGMENT

The work reported in this paper has formed a part of the IoN Work Area of the Core 3 Research Programme of the Virtual Center of Excellence in Mobile & Personal Communications, Mobile VCE, www.mobilevce.com, whose funding support, including that of EPSRC, is gratefully acknowledged. Fully detailed technical reports on this research are available to Industrial Members of Mobile VCE.

REFERENCES


Karo Ann Chew received his B.Eng. degree in electronics and electrical engineering and Ph.D. degree in mobile communications in 1998 and 2004, respectively, from the University of Surrey, Guildford, U.K.

He was a Research Fellow in Mobile Communications Research Group, Center for Communication Systems Research (CCSR) in University of Surrey between 2000 and 2004. He worked in and managed several research projects funded by the Mobile VCE, U.K. and the EU in the areas related to All-IP mobile network, mobility management, ad hoc routing, resource management and the interworking of digital broadcasting and mobile networks. Since 2004, he has been a Senior Research Professional in the Mobility Research Centre of British Telecommunications, U.K., working in various areas related to mobile and wireless communications and fixed-mobile convergence.

Dr. Chew is the recipient of the 2003 Vodafone Prize for Research Excellence. He is a Member of IEE and IEE U.K.

Shyamalie Thilakawardana received her B.Sc. degree in computer science and engineering from the University of Moratuwa, Moratuwa, Sri Lanka, and the M.Sc. degree in telecommunication and software engineering and the Ph.D. degree in mobile communications, in 1998 and 2002, respectively, from the University of Surrey, Guildford, U.K.

Since 2003 she has been a Post Doctoral Research Fellow in Mobile Communications Research Group, Center for Communication Systems Research (CCSR), University of Surrey. Her research interests have focused on resource scheduling, spectrum management and cognitive radio. She has authored and co-authored numerous research papers in referred journals and international conferences. She holds a U.K. patent in the area of radio resource management.

Dr. Thilakawardana is the recipient of the 2002 INmasat Prize for Research Excellence. She is a Member of IEE U.K.

Yajian Liu (S’05) received the B.Eng. and M.Eng. degrees from the School of Electrical and Information Engineering, Hunan University, China, in 1998 and 2001, respectively. He is currently working toward the Ph.D. degree at the Center for Communication Systems Research (CCSR), University of Surrey, U.K.

From 2001 to 2003, he was a full-time researcher on mobile communication systems in Shanghai R&D Center of Huawei Technologies Co. Ltd., China. Since 2003, he was actively involved in IST European projects MUMOR, WINNER, and FIREWORKS. His research interests include radio resource management for multi-hop cellular networks, cooperative relaying for OFDM/OFDMA systems, and packet scheduling for 3G and beyond systems.

Klaus Moessner is a senior research fellow for Mobile Communications in the Centre for Communication Systems Research at the University of Surrey, U.K. Klaus earned his Dipl-Ing (FH) at The University of Applied Science in Offenburg, Germany, an M.Sc. from Brunel University, U.K. and his Ph.D. from the University of Surrey (U.K.). Klaus’ research interests include reconfigurability of the different system levels, including reconfiguration management and scheduling in wireless networks and adaptability of multimodal user interfaces. He also works in the area of resource efficiency and mechanisms for dynamic resource allocation.

Authorized licensed use limited to: University of Surrey. Downloaded on February 18,2010 at 11:59:05 EST from IEEE Xplore. Restrictions apply.
Rahim Tafazolli (M’89) is a Professor of Mobile/Personal communications and Head of Mobile Communications Research at Center for Communication Systems Research (CCSR), University of Surrey, U.K.

He has been active in research for over 20 years and has authored and co-authored more than 300 papers in refereed international journals and conferences.

Professor Tafazolli is consultant to many mobile companies, has lectured, Chaired and has been invited as keynote speaker to a number of IEE and IEEE workshops and conferences. He has been technical advisor to many mobile companies, and European Union all in the field of mobile/wireless communications. He is the Founder and past Chairman of IEE International Conference on 3rd Generation Mobile Communications. He is member of Motorola Visionary Board, Chairman of EU Expert Group on Mobile Technology Platform, past chairman of WG 3 of WWRF, and editor of the books “Technologies for the Wireless Future” Vol. 1 2004 and Vol. 2 2006. He is nationally and internationally known in the field of mobile communications and acts as external examiner for the British Telecom MSc Course.