

QoS and Channel Aware Packet Bundling for VoIP and Data Traffic in Multi-Carrier Cellular Networks

Cory Beard, Baek-Young Choi, and Hyungbae Park
University of Missouri-Kansas City, Kansas City, Missouri 64110 USA
Email: beardc,choiby,hpark@umkc.edu

Abstract—We study the problem of multiple packet bundling in multi-carrier cellular networks for VoIP and data traffic. In a time-slotted system such as the cdma2000 1xEV-DO downlink, part of a time-slot may not be fully utilized, if the packet sizes are small, as in the case of real-time VoIP traffic. Packet bundling can alleviate such a problem by sharing a time slot among multiple users, as proposed in the EV-DO Revision A system. A recent revision, EV-DO Revision B, enables further increased system capacity through multiple carriers. However, the efficacy of packet bundling, especially across multiple carriers, is not well understood. When packets are bundled, the packet with the lowest channel quality dictates the modulation and coding format of the entire bundle, possibly wasting significant slot space due to unnecessary coding bits for the other packets. We first present how multiple carriers, together with packet bundling, improve spectral efficiency, but lose some of that benefit in realistic channel scenarios. Then we propose an effective heuristic algorithm to effectively exploit the multi-carrier with bundling system. We consider QoS requirements as well as time-varying and user-dependent channel conditions, and then show how channel utilization can be significantly improved while keeping delays of real-time packets limited.

I. INTRODUCTION

Third and fourth generation wireless systems promise substantial capacity improvements over previous systems through their support of VoIP and data traffic through various techniques including rapid channel quality indication feedback, adaptive modulation and coding, multiple carriers, hybrid ARQ, and predictive and soft handover.

Several of these systems use time-slotted mechanisms for sending packets. For VoIP packets, however, the packets are frequently much smaller than the capacity of the time slot. Instead of wasting large parts of time slots to send small VoIP packets, some systems allow the use of multiuser packets (MUPs) so that multiple VoIP packets take full advantage of a time slot, along with also being able to include data packets. Not all mobiles have the same channel quality between sender and receiver, however. In such cases, those packets with the *worst* quality channel in a selected group of packets will dictate the capacity of a particular time slot so that all packets can be successfully received. If packet bundling is not executed according to an efficient algorithm, much of the benefits of multiuser packets would go unrealized.

This work was supported in part by the US National Science Foundation under Grant No. 0729197. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the US National Science Foundation.

Most recently, standards like EV-DO Rev. B have been developed to also include multiple carriers along with multiple packets per time slot. If an environment is frequency selective, the multiple carriers to a user will have different small-scale multipath fading characteristics but similar large-scale and shadowing attributes. Scheduling algorithms must be aware of and take advantage of this fact for the addition of multiple carriers to also reach its full potential.

This paper provides several important contributions to making multicarrier MUP systems successful. In this work, we accomplish the following: i) present analytic models for the multicarrier bundling system, both for ideal and realistic channels with small-scale fading. We show theoretically that the addition of both multiple carriers and multiple packets per slot has more advantages than trying to use either approach by itself with more carriers or more packets per bundle. ii) demonstrate that realistic channel conditions severely restrict the benefits that could be gained from bundled MUPs. In one example, approximately 40% less throughput is achieved in a realistic channel. It is important to develop scheduling algorithms that shrink this 40% gap. iii) show that the problem of optimal packet bundling in a multi-carrier system is NP-complete, and thus develop a simple yet effective heuristic algorithm to serve both real-time VoIP and best effort data traffic, showing channel utilization to be maximized or controlled while protecting QoS of VoIP traffic. iv) perform extensive simulations to show the efficacy of the heuristic algorithms and impacts of the parameters.

The rest of the paper is organized as follows. We discuss the related work regarding wireless scheduling in Section II. Section III gives background discussion to set the context for our study of the multi-carrier bundled wireless system. In Section IV a multicarrier multiuser analytical model is presented for ideal and realistic wireless channels. Then Section V follows with a discussion of candidate algorithms for optimizing QoS, throughput, and fairness while maintaining system stability. Section VI provides extensive simulation results and comparisons of the algorithms. We present conclusions in Section VII.

II. RELATED WORK

In recent years, research and development efforts have increased on adaptive wireless systems where higher rate and power levels are allocated as the channel quality increases.

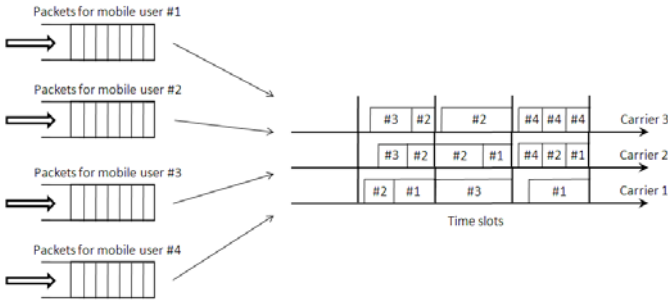


Fig. 1. The concept of packet bundling.

This enables physical layer Adaptive Modulation and Coding (AMC). Relying on AMC, opportunistic schedulers select the user with the best channel quality to maximize the channel utilization. However, QoS may be violated for some users in such schemes. The work in [12] shows that Delay-Margin-based Scheduling nested with User-Channel-based Scheduling performs well both in delay and utilization metrics.

Multi-carrier scheduling research has received attention recently. Slot-by-slot multi-carrier algorithms were first proposed in [3]. Frame-based multi-carrier scheduling is investigated in [4] where a scheduling decision is made in a batch of a frame, and it attempts to spread out packets of different users within a template so that delay jitter would be reduced. Multi-carrier scheduling variants with contiguity constraints are studied in [5]. Those studies assume that only one user can be scheduled for a carrier at a time slot, and all packets are treated equally for a common objective.

Studies on EV-DO VoIP capacity are presented in [6]. [20] shows the trade-off between delay and system throughput in terms of the number of users, with opportunistic scheduling using an analysis and simulation of the EV-DO system. The authors in [7] developed a soft algorithm that has an additional step for VoIP packets in order to check the channel condition, that is, whether the current data rate is larger than or equal to the average data rate. They demonstrated that Proportional Fair (PF) scheduling combined with the soft PF algorithm shows the best performance over Max-Rate algorithms. QoS and channel aware packet bundling algorithms are proposed in [15] for single carrier EV-DO networks. To the best of our knowledge, there has been little work on packet bundling for VoIP and data traffic on multi-carrier systems. This work extends [15] by providing multi-carrier algorithms and analytical models.

III. BACKGROUND

In this section, we present background discussion to set the context for our study of the multi-carrier bundled wireless system with the example of EV-DO networks.

In EV-DO networks, the downlink channel is time slotted and shared by all users in a cell [9], [10]. Early systems allowed one selected user to receive data in a single time slot; later systems allow multiple users to receive packets during a time slot.

Signal strength in a wireless system is location dependent (i.e., user dependent) due to large-scale fading based on the distance from the base station and shadowing from obstructions. Signal strength is also time varying due to multipath and Doppler spread (movement of mobiles). This combination of slow fading and fast fading, along with interference from other signals, results in degradation of the Signal to Interference-plus-Noise Ratio (SINR) [22]. A high SINR yields a high data rate and low error, but a lower SINR can be overcome if the base station (BS) estimates each user's channel condition and uses Adaptive Modulation and Coding (AMC) to improve bit error rates. AMC also causes a somewhat lower throughput, however, since more coding bits are added and lower-order modulation schemes are used.

The BS receives feedback from individual mobile station (MS) channel measurements through channel quality indication (CQI) feedback. AMC schemes are adopted in many currently deployed and future wireless standards, including cdma2000 1xEV-DO, WCDMA, IEEE 802.16 broadband wireless access (WiMax), and the IEEE 802.11 Wireless LAN. In EV-DO, the measured CQI value is estimated by the MS through packet loss rates and is fed back to the base station once every 1.667 msec using a reverse control channel. This time duration is short enough so that each user's channel quality stays approximately constant within one time slot.

In time slotted systems, the number of users supported is limited theoretically by the number of time slots per second and packet arrival rates (i.e., $\frac{no_time_slots/sec}{packet_arrival_rate/user}$).¹

The multi-user packet (MUP) was introduced in EV-DO Rev. A. In addition to supporting more users per given time period, delay deadlines of real-time applications like VoIP can be better met with multi-user packets. The VoIP application is a good fit for the multi-user packet, since VoIP packets are generated regularly and frequently (every 20 ms when active) and their sizes are small. MUPs can also increase the throughput of best-effort (BE) traffic, since VoIP and BE traffic can share the same frame. A bundled packet is recognized from the preamble of the physical layer packet and the MAC header.

A major limitation to the effectiveness of the multiuser packet relates to Adaptive Modulation and Coding. Each MS that is to receive a packet in the current MUP could have a different channel condition. All other packets in the bundle will be encoded with the worst AMC format, even if the extra coding is not really needed. Also, throughput is not the only consideration since delay is a critical concern for VoIP packets. In this study, we develop a MUP bundling algorithm to balance throughput and delay, since they are somewhat competing objectives. We also consider fairness. Users with poor channel quality will likely stay that way since they are probably far from a base station, but they need throughput and

¹In fact, the total number of supported users can also be limited by the reverse link. In EV-DO, the reverse link uses CDMA, so multiple MSs send transmissions concurrently and the EV-DO system capacity is limited by the interference level measured by RoT (Rise over Thermal). However, we focus our study on the forward link only.

QoS as well.

Multi-carrier systems also provide benefits to add to the advantages of packet bundling. In terms of small scale fading, multi-carrier systems usually operate in frequency selective conditions, so the multiple carriers will have different small-scale fading characteristics. A user that might be in a deep fade on one carrier could have a good signal on another carrier. Packet bundling should take advantage of that.

IV. ANALYTICAL MODEL

This section provides several analytical tools for understanding a multi-carrier multi-user packet system. It also shows how to choose an effective scheduling discipline to take advantage of the potential benefits of multiple carriers and multiple packets per time slot when the effects of realistic channel variations are taken into account.

A. Model for an ideal channel

We first consider how to model an ideal channel serving multiple packets per time slot, then add the multi-carrier characteristic. We initially assume a *bulk service system* based on the Markov model of an ideal channel ([16], [18]). In a bulk service system, arrivals occur individually, but service to those packets is done in a bulk manner, where b packets are processed at the same time at a rate μ and finish together. If the arrivals are assumed to be Markov, and the service times are also Markov, the system can be described by an $E_r/M/1$ system where the E_r notation indicates an r -stage Erlangian arrival process. We do actually believe that real-world traffic or service times are Markov in the algorithms we develop, but just use these assumptions so we can have a model to show comparisons between single user, single carrier systems and multi-user, multi-carrier systems. Matrix exponential methods [17] [19] could readily be applied to extend the models to more complex, realistic arrival and service processes.

We extend this model further by including multiple wireless channels, both ideal and realistic channels where channel conditions limit bundling capacity. Related recent work on multiserver bulk service systems includes analysis of discrete-time queues [11] and approximations for multiserver bulk systems [13]. In our multi-carrier bulk service case, systems can be viewed as $E_r/M/C$ systems where service can be provided by any of C available identical carriers.

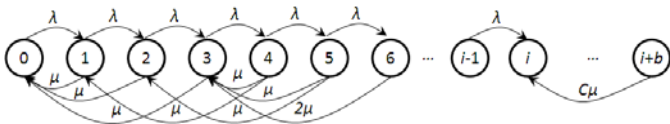


Fig. 2. Markov chain for multi-carrier bulk service system.

The state diagram for such a system is shown in Fig. 2, where each state indicates the number of packets in the system. When there are less than b packets in the system, a service rate of μ is shown by an arrow that serves all available packets and goes to state 0. When there are more than b packets in

the system, but less than bC packets, multiple transitions will come out of a state to indicate either full batches of size b or smaller batches from the remaining packets in the state. When 2 full batches come from a state, this means the transition from state i to $i - b$ is indicated with a rate of 2μ . Finally, once the state number is larger than bC , then coming out of the state will be batches of b packets at a total rate of $C\mu$. The queue size for now is assumed to be infinite. To solve the system, one needs to define and solve the balance equations for flows coming in and out of each state. Results that compare different numbers of carriers and batch sizes are provided after the derivations in the next subsection.

B. Model for a realistic channel

In the analytical model presented so far, a perfect channel was assumed; whenever there are b packets in a queue and a channel is available, send the b packets together in a bundle. All channels have this ideal characteristic.

A realistic channel model, however, would consider short-term fading, large-scale pathloss, and frequency selective fading, some of which we add to our analytical model. Packets cannot always be bundled together with a maximum batch size and still fit in the time slot because on a slot-by-slot basis some users will have poor channel conditions that require larger packets. Frequency selective fading will also cause channel conditions to be different across carriers.

To incorporate short-term fading, we assume speech is encoded using a variable rate vocoder via the Enhanced Variable Rate Codec (EVRC) that is recommended for EV-DO. EVRC creates packets of 171, 80, 40, or 16 bit packets, depending on the voice activity. For our model we assume all VoIP packets are 80 bits. EV-DO Rev. A has 14 packet formats that allow 1, 2, 4, 8, 8, 16, 16, 24, 32, 32, 48, 64, 40, or 80 64-bit packets per slot, depending on the channel quality to the user as measured by the DRC (data rate control) sent from the mobile to the base station. This means that there can be 64, 128, 256, 512, 512, 1024, 1024, 1536, 2048, 2048, 3072, 4096, 2560, or 5120 bits per slot. Since 8 80-bit EVRC packets take 640 bits, there are 9 out of the 14 formats that could support a EV-DO's maximum bundle size of 8. Two formats could support 6 packets, one could support three, and the last two could only support about one packet. We assume that each DRC level can occur and is equally likely. At each time slot, the DRC can be completely different. This models small-scale fading, but not large scale fading. Future work will also include large-scale considerations. The probabilities for each of 4 channel capacities are $p_8 = 9/14$ (9 out of 14 formats can support 8 packets), $p_6 = 2/14$, $p_3 = 1/14$, and $p_1 = 2/14$.

It is not so simple as to just use these values, however. We must also select a group of packets to bundle. This is the main focus of this paper. If we select a group and one of them has a channel quality that can only support 3 packets, then even if the others could have supported more, the current slot can only support 3 packets, since the most robust modulation and coding must be used to cover the poorest quality channel.

For this analytical model, we introduce QoS-aware Multi-carrier Packet Bundling (QMB). One approach to QMB takes the packet with the longest delay and bundles it with the other packets with the next longest delays so they can be sent before they violate their deadlines. To approximate this here, choose to send the head-of-line packet in the queue (most likely the one with the longest delay), then take the next packets after it to be bundled. To send 8 packets together, all of the first 8 packets in the queue must have a DRC that allows for 8 packets. If in the process of building the bundle the next packet has a DRC that cannot be accommodated, then stop, bundle, and send the packets that have been chosen so far. The probabilities of sending each bulk size, therefore, are as follows.

- Bundle of 8 packets: $p_8^8 = 0.029$
- Bundle of 6 packets: $(1 - p_1 - p_3)^6 (1 - (1 - \frac{p_6}{p_6 + p_8})^6) + (p_6^6)(1 - (1 - p_1 - p_3 - p_6)^2) = 0.206$
- Bundle of 3 packets: $(1 - p_1)^3 (1 - (1 - \frac{p_3}{p_3 + p_6 + p_8})^3) + (1 - (1 - p_1 - p_3)^3)(1 - p_1 - p_3)^3 = 0.395$
- Bundle of 1 packet: $1 - (1 - p_1)^3 = 0.370$

These values are incorporated into the model by having multiple possible transitions from a state where there used to be only one at the maximum batchsize. For example, if state i previously could allow a bundle of 8 packets, with a transition rate of μ back to state $i - 8$, now it would have transitions of 0.029μ to state $i - 8$, 0.206μ to $i - 6$, 0.395μ to $i - 3$ and 0.370μ to $i - 1$. If state j only allowed a bundle of up to 4 packets, now it would have 0.370μ to $j - 1$, 0.395μ to $j - 3$, and $(0.029 + 0.206)\mu = 0.235\mu$ (6 or 8 allowed) to $j - 4$.

C. Results and comparisons

Fig. 3 shows the performance of the ideal channel and the realistic channel as the batch size (b) grows for 1 channel. We use $\mu = 600$ to emulate EV-DO's 600 slots/sec. The case of $b = 1$ matches the $M/M/1$ result of $T = 1/(\mu - \lambda) = 1/(600 - 550) = 20$ msec. Simply using a batch size of 2 creates a substantial reduction in delay. Even though EV-DO supports up to 8 packets per bundle, the main benefit is seen for up to 4 packets. For the ideal channel as b increases, the system approaches the result for an $M/M/\infty$ case of $T = 1/\mu = 1/600 = 1.67$ msec. As b increases, the result for the realistic channel approaches 2.33 msec., so the realistic channel experiences approximately 0.67 msec. more delay at these arrival and service rates.

The effects of having multiple channels are seen in Fig. 4, with a batch size of 1. Once again, the single channel result matches the $M/M/1$ result and many channels produce the $M/M/\infty$ result. Just having two channels produces a large benefit and more than 3 channels provides no appreciable additional benefit at these arrival and service rates. Only one curve is shown on this plot because the realistic channel model has no difference in performance between channels.

A combined picture of multi-carrier and multi-user packets is provided in Fig. 5 for the ideal channel case. The realistic channel case produces a similar plot, so is omitted here. One

TABLE I
AVERAGE DELAY FOR AN IDEAL CHANNEL

	$b=1$	$b=2$	$b=3$	$b=4$
$C=1$	19.9697	2.5121	1.9302	1.7718
$C=2$	2.1099	1.6962	1.6701	1.6671
$C=3$	1.7253	1.6677	1.6667	1.6667
$C=4$	1.6749	1.6667	1.6667	1.6667

can see that, compared to the $M/M/1$ case, adding another channel has more benefit than increasing the batch size to 2. This can also be seen in Table I where delay goes down to 2.1099 msec. instead of 2.5121 msec. Note also that going to 3 channels is not as good as taking two channels and adding a multi-user packet with batch size of 2. One has to go to a batch size of 5 in the single carrier case to improve upon the (2, 2) case.

The final figure for this analytical analysis is probably the most important and enlightening. In Fig. 6, the delay is shown for increasing values of λ . The system has a single carrier and a batch size $b = 4$. Both ideal and realistic channel models are shown. This plot can be used to determine what might be considered to be the capacity of this system. One measure might be the place where each curve crosses a threshold, say an average delay of 4 msec. In that case, the ideal channel can handle up to $\lambda \approx 1700$ but only up to $\lambda \approx 900$ for the realistic channel. The ideal channel has $\approx 47\%$ less capacity. If one estimates where a system "blows up", then we have $\lambda \approx 1300$ for ideal and $\lambda \approx 2300$ for realistic, $\approx 44\%$ reduced capacity. The realistic model used QMB, which emphasized QoS over capacity, so we will see later in the paper that other scheduling approaches seek to get closer to the ideal capacity. But the message is clear from this figure; realistic scenarios have a substantial decrease in capacity from the ideal and it is important to find the best scheduling scheme possible. The benefits to finding such new schemes are substantial.

V. MULTI-CARRIER PACKET BUNDLING ALGORITHMS

Now we shift to describing algorithms that can be used to exploit multi-carrier packet bundling. In a single carrier, single traffic type, and no bundling system, Max-Weight (MW) and Proportional Fair (PF) are two popular kinds of scheduling algorithms. In MW, a user u that maximizes the value of $Q(u)R(u)$, is chosen for scheduling, where $Q(u)$ is the queue size of user u , and $R(u)$ is the channel rate of user u . PF selects a user u , which maximizes the ratio of $R(u)/\mu(u)$, where $\mu(u)$ is the long term average rate of a user u . MW is suitable in a small or finite buffer system, as it has the stability property, where queue size is kept bounded whenever possible. In a large or infinite buffer system, the PF algorithm is desirable since it is known to maximize the utilization when the system is stationary.

In this study, we extend the above objectives in two ways, to a multi-carrier packet bundling problem and to one that serves both real-time and data traffic. For simplicity of our discussion, we focus our algorithms on the MW type in serving BE traffic,

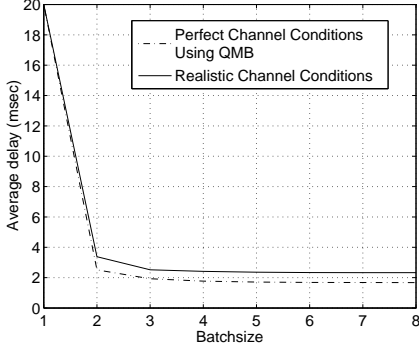


Fig. 3. Avg. delay versus batch size: 1 channel, $\lambda = 550$, $\mu = 600$.

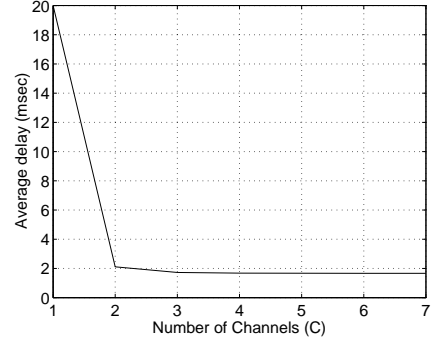


Fig. 4. Average delay versus number of channels: batchsize = 1, $\mu = 600$, $\lambda = 550$.

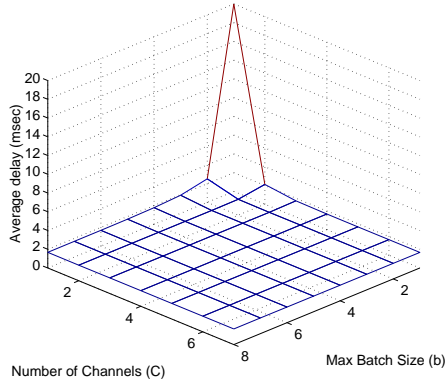


Fig. 5. Multi-carrier and multi-user performance for ideal channels: $\mu = 600$, $\lambda = 550$.

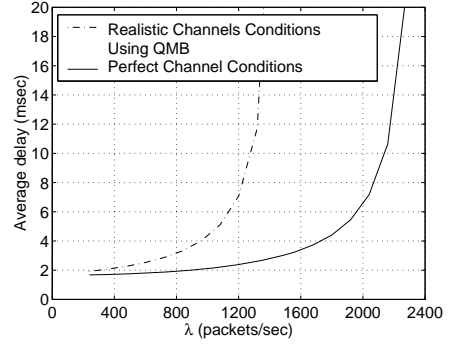


Fig. 6. Multi-carrier multi-user performance for realistic channels: Batchsize = 4, $\mu = 600$ packets/sec., $C = 1$.

but can be modified to the PF algorithm easily. The scheduling decision is made slot-by-slot, and we omit the time index in the following discussion. We also assume BE packets are always available for transmission, and a BE packet is large enough to fill an entire time slot. Then, our objective is for a given time slot, to find a set of packets to be served, B that satisfies the following:

$$\max_{p(u) \in B} \sum_c \sum_u Q(u) R(u, c) \quad (1)$$

such that

$$d(p_u) \leq D_{thresh}$$

Where $R(u, c)$ is the channel rate of c for user u , p_u is the head-of-line packet of user u , $d(p_u)$ is the delay of p_u , and D_{thresh} is the delay requirement. Note that we use u for either a VoIP user or a BE user in the following discussions; it can be also used for a *flow* when a user has both VoIP and BE traffic.

The above problem is a complex problem, where sub-problems can be shown to be NP-complete. Finding packet assignments for a multi-carrier system without bundling is shown as NP-complete [3], and the problem of finding a set of packets for a given channel is also NP-complete as

shown in the Appendix. Thus, we develop heuristic algorithms that approximately optimize QoS requirements, utilization, and both QoS and channel utilization, respectively.

Algorithm 1 QMB

```

if VoIPSet  $\neq \emptyset$ 
    put the longest delayed VoIP
    for its best carrier  $c$ , set  $R(c) := R(u, c)$ 
while VoIPSet  $\neq \emptyset$  and AvailChSet  $\neq \emptyset$ 
    put the next longest delayed VoIP
     $R(c) = \min_u \{R(u, c), R(c)\}$ 
foreach carrier  $c$  not assigned yet
    put a BE packet of  $u$  with  $\max_u R(u, c)$ 

```

A. QoS Aware Multi-Carrier Packet Bundling (QMB)

We first present a scheduling algorithm that mainly considers a QoS requirement, and performs bundling in the order of QoS requirement. With this so-called QoS aware multi-carrier packet bundling (QMB), VoIP packets are first considered in the decreasing order of delay. We discuss QoS mainly in the context of delay parameter. However, it can be extended to other QoS parameters. The algorithm works as follows. First,

Algorithm 2 CMB

foreach carrier c
 if $VoIPSet \neq \emptyset$
 find the most common $R(c)$ of VoIP users
 if the number of VoIP of $R(c) \geq B_{thresh}$
 while $VoIPSet \neq \emptyset$ and T is not full
 add a VoIP of the $R(c, u)$ until T is full
 if T is not full
 put a BE of u with $\max_u Q(u)R(u, c)$
 else
 Add a BE packet of highest MW into T ;

Algorithm 3 QCMB

while delay of a VoIP $\geq D_{thresh}$
 run QMB algorithm
foreach carrier c not assigned yet
 run CMB algorithm

a packet with the longest delay, p_u^{*d} will be selected for a service from the VoIP queue.

$$p_u^{*d} = \arg \max_u d(p_u) \quad (2)$$

where $d(p_u)$ is the delay of a packet of user u .

Then the highest quality channel of the user, $c^* = \arg \max_c R(u, c)$ is assigned for the packet. The next longest delayed VoIP packets are then de-queued to be bundled together with the assigned packet on the carrier with $R(c) = \min_u R(u, c)$, while the time slot of interval T has available room for more VoIP packets. If there is no other VoIP to be bundled together and the time slot is still available, then BE traffic will be included in the bundle. The sketch of this QMB algorithm is shown in 1.

B. Channel Aware Multi-Carrier Packet Bundling (CMB)

Channel aware multi-carrier packet bundling (CMB) runs carrier-by-carrier, and *packets from MSs with the same channel rate* are bundled together in order to better utilize the channel. It also attempts to do multiple packet bundling only when there are enough VoIP packets of the same channel quality. Otherwise, the time slot can be used for a big BE packet so that the channel is maximally utilized. The high bundling ratio of the packets of the same channel rate enables efficient channel utilization.

A bundling threshold, B_{thresh} , is defined, which is the minimum real-time data size or the minimum number of VoIP packets that must be bundled. Large B_{thresh} forces the real-time packets to be bundled with a high bundling ratio, in order to better utilize the channel with BE traffic. When B_{thresh} is set to be small, VoIP packets can be scheduled without having to be deferred much. Small B_{thresh} may be used when there are not many VoIP packets to be bundled. Note that since the objective is only to maximize the utilization, CMB does not provide any delay guarantees. Thus, a packet may wait

for a long time for a chance of bundling. Real-time packets that exceed the maximum allowed delay, or packets arriving when the queue is full, will be dropped. The CMB algorithm is depicted in Algorithm 2.

C. QoS and Channel Aware Multi-Carrier Packet Bundling (QCMB)

Since our goal is to improve the channel utilization while satisfying the QoS constraints, we now jointly consider QoS and channel condition to provide an algorithm that capitalizes on the benefits of both QMB and CMB. Thus the packet bundling is performed only within the QoS budget of real-time packets. We use the delay requirement D_{thresh} as the maximum allowed delay; this allows the scheduling of real-time packets to be *deferred* in the queue without violating QoS. If there are packets whose delays are greater than or equal to D_{thresh} , those packets should be bundled first in order to meet the delay requirement. Otherwise, QCMB attempts to utilize the channel efficiently by gathering packets of similar channel conditions that can be bundled together. The impact of the control parameter D_{thresh} is investigated in detail in the next section. The QCMB algorithm is outlined in Algorithm 3.

VI. EVALUATION

As the effectiveness of the bundling algorithms would depend on the traffic mix, we evaluate the algorithms under various scenarios using our own simulator. We have a single base station and vary the number of VoIP sessions from 10 to 80 users. Additionally, 10 Best Effort (BE) sessions are added to observe the interplay of VoIP and BE traffic. For BE traffic, FTP file downloads are performed for large files, so that the channels would never go idle.

We first overview the queue sizes of the three algorithms in Figures 7, 8, and 9. All of the algorithms show stable queue sizes over time, and QMB shows the smallest VoIP queue sizes that indicates the lowest delays. The largest queue sizes of CMB are caused by serving BE traffic which we assume is always present to fill up the available bandwidth. QCMB exhibits the middle of the two algorithm performances.

Figures 10, 11, and 12 present the average VoIP delay and BE throughput over different numbers of VoIP users. VoIP delay for CMB and QCMB at times goes down with more load because more VoIP packets can be bundled.

The three algorithms are directly compared in terms of BE throughput and VoIP delay in Figures 13 and 14, respectively. QCMB displays the performance tradeoff between QMB and CMB. When there are small number of VoIP users in the system, QCMB behaves close to CMB. However, as the number of VoIP users increases, QCMB tends to work like QMB.

Next we investigate the impact of algorithm parameters, B_{thresh} and D_{thresh} , in Figures 15 and 16, respectively. B_{thresh} defines how many VoIP packets we should wait for before we bundle and transfer VoIP packets. Thus, if it is $B_{thresh} = 1$, the scheduler will send VoIP packets right away without waiting for the next VoIP packets coming. This

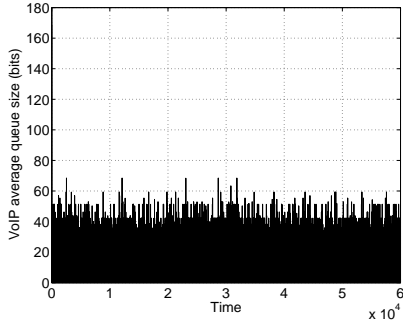


Fig. 7. Average queue size of QMB: VoIP users = 10.

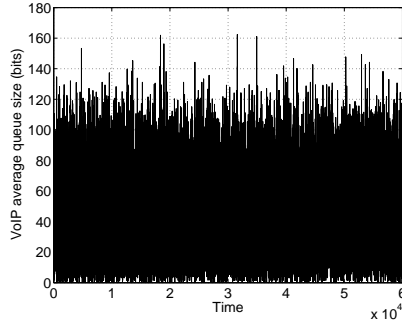


Fig. 8. Average queue size of CMB: VoIP users = 10.

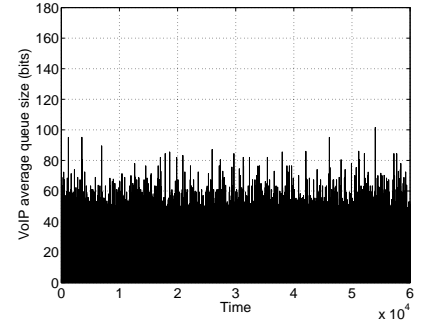


Fig. 9. Average queue size of QCMB: VoIP users = 10.

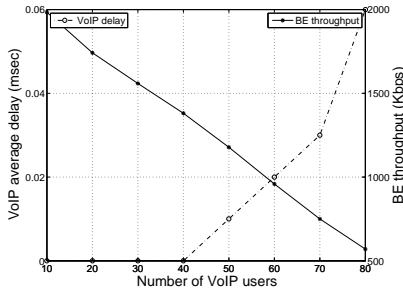


Fig. 10. Average VoIP delay and BE throughput of QMB.

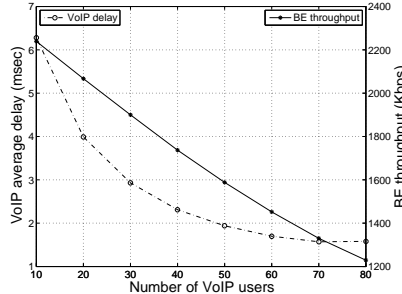


Fig. 11. Average VoIP delay and BE throughput of CMB.

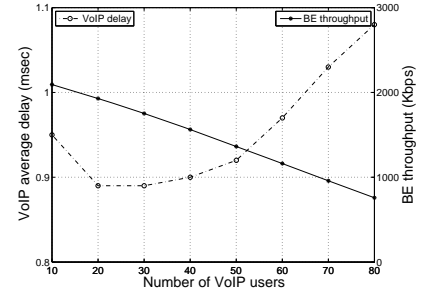


Fig. 12. Average VoIP delay and BE throughput of QCMB.

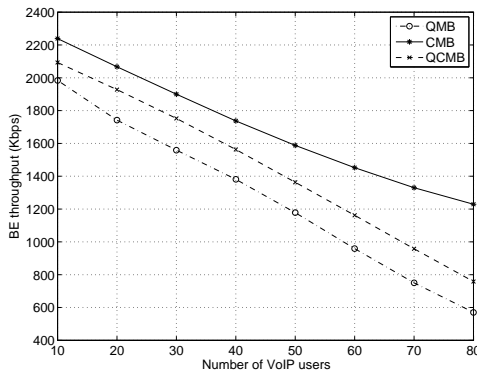


Fig. 13. Comparison of BE throughput: $D_{thresh} = 1, B_{thresh} = 4$.

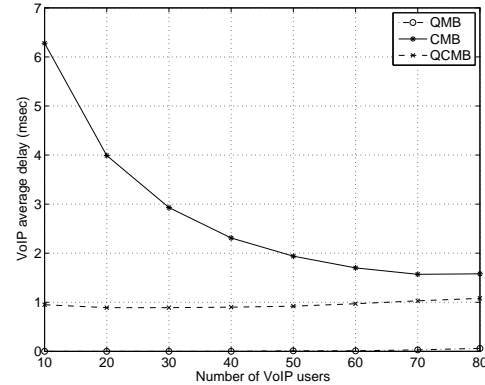


Fig. 14. Comparison of delay: $D_{thresh} = 1, B_{thresh} = 4$.

strategy is good for performance of VoIP delays but not for BE throughput. When D_{thresh} is small, no matter the B_{thresh} value, QCMB will show the lowest delays since QCMB works like QMB. That's why Figure 16 has the smallest delay for $D_{thresh} = 1$. However, it wastes channel utilization. As the D_{thresh} increases, probabilities satisfying D_{thresh} get lower, which causes QCMB behave close to CMB. The evaluations show that QCMB the improves system utilization (throughput) while keeping the QoS requirement (small delay), and achieves a good tradeoff between QMB and CMB.

VII. CONCLUSIONS

We investigated the problem of multiple packet bundling for multi-carrier cellular networks that serve both real-time VoIP

and data traffic. We have developed analytical models and have shown that multiple carriers and packet bundling together provide higher channel capacity than individual technologies with a higher level of multi-carrier or packet bundling degree. Some of that benefit, however, is lost in realistic channel scenarios. As a matter of fact, capacity can be 50% less in realistic scenarios. We also found that optimal scheduling for packet bundling in multi-carrier systems is an NP-complete problem, and thus proposed three heuristic algorithms, namely QMB, CMB, and QCMB. The superior QCMB algorithm considers QoS requirements of VoIP traffic as well as time varying and user dependent channel conditions. Through various simulations, we have demonstrated that channel utilization can be significantly improved while keeping delays of real-

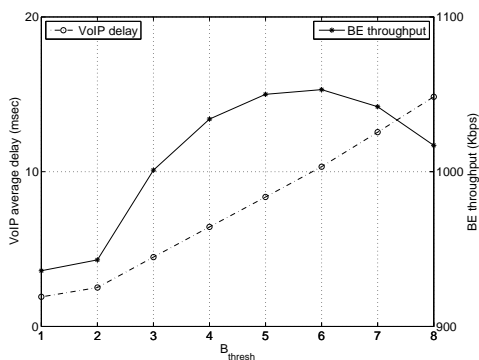


Fig. 15. Impact of B_{thresh} in CMB: VoIP users = 10.

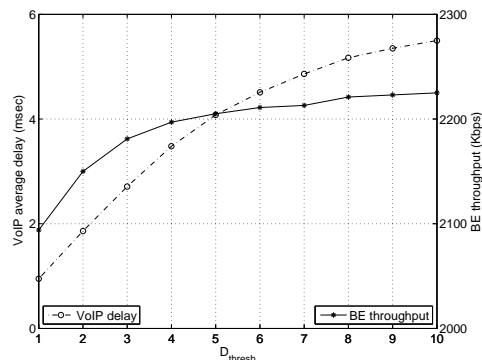


Fig. 16. Impact of D_{thresh} in QCMB: VoIP users = 10, $B_{\text{thresh}} = 4$.

time packets limited. Particularly, QCMB provides a good trade-off between small VoIP delay and high BE throughput.

To the best of our knowledge, this is the first work that addresses the multi-carrier packet bundling problem. We believe our study provides important insights on designing multi-carrier and bundling systems and for devising practical schemes that realize high spectral efficiency in wireless networks that support both VoIP and BE traffic.

REFERENCES

- [1] TIA 45.5/98.04.03.03. The cdma2000 ITU-R RTT Candidate Submission, April 1998.
- [2] M. S. Alouini and A. J. Goldsmith. Adaptive modulation over Nakagami fading channels. *Kluwer Journal of Wireless Communication*, 13(1–2):119–143, May 2000.
- [3] Matthew Andrews and Lisa Zhang. Scheduling algorithms for multi-carrier wireless data systems. In *ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2007.
- [4] Matthew Andrews and Lisa Zhang. Creating templates to achieve low delay in multi-carrier frame-based wireless data systems. In *Proc. IEEE Infocom*, 2008.
- [5] Matthew Andrews and Lisa Zhang. Contiguous-carrier scheduling algorithms for multi-carrier wireless systems. In *Proc. IEEE Infocom*, 2009.
- [6] Qi Bi, Pi-Chun Chen, Yang Yang, and Qinqing Zhang. An Analysis of VoIP Service Using 1 EV-DO Revision A System. *IEEE Journal On Selected Areas in Communications*, 24(1):36–45, 2006.
- [7] Young-Jun Choi and Saewoong Bahk. Channel-aware VoIP packet scheduling in cdma2000 1x EV-DO networks. *Elsevier Journal of Computer Communications*, 30:2284–2290, 2007.
- [8] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman, 1979.
- [9] Vijay K. Garg. *CDMA IS-95 and cdma2000*. Prentice Hall, 2000.
- [10] Vijay K. Garg. *Wireless Communications and Networking*. Morgan Kaufmann Publishers, 2007.
- [11] V. Goswami and G.B. Mund. Multiserver bulk service discrete-time queue with finite buffer and renewal input. *Computers and Mathematics with Applications*, 57:1377–1388, 2009.
- [12] Quang-Dung Ho, Mohamed Ashour, and Tho Le-Ngoc. Channel and Delay Margin Aware Bandwidth Allocation for Future Generation Wireless Networks. In *Proc. IEEE Globecom*, New Orleans, LA, Nov 2008.
- [13] Ming-Guang Huang, Pao-Long Chang, and Ying-Chyi Chou. Analytic approximations for multiserver batch-service workstations with multiple process recipes in semiconductor wafer fabrication. *IEEE Transactions on Semiconductor Manufacturing*, 14:395–405, 2001.
- [14] TIA IS-127. Enhance Variable Rate Codec (EVRC) 8.5 kbps Speech Coder.
- [15] Jung Hwan Kim, Baek-Young Choi, and Cory Beard. Qos and channel aware packet bundling for voip traffic in cellular networks. In *International Teletraffic Congress*, 2009.

- [16] Leonard Kleinrock. *Queueing Systems Volume I: Theory*. John Wiley and Sons, 1975.
- [17] L. Lipsky. *Queueing Theory: A Linear Algebraic Approach*. New York:MacMillan, 1992.
- [18] Marcel F. Neuts and R. Nadarajan. A multiserver queue with thresholds for the acceptance of customers into service. *Operations Research*, 30:948–960, 1982.
- [19] Zhefu Shi, Cory Beard, and Ken Mitchell. Analytical models for understanding misbehavior and mac friendliness in csma networks. *Performance Evaluation*, 66:469–487, September 2009.
- [20] Roshni Srinivasan. *Scheduling in Packet Switched Cellular Wireless Systems*. PhD thesis, University of Maryland, College Park, 2004.
- [21] 3GPP2 C.S0024-0 v2.0. cdma2000 High Rate Packet Data Air Interface Specification. http://www.3gpp2.org/public_html/specs/C.S0024_v2.0.pdf, Oct. 2000.
- [22] B. H. Walke. *Mobile Radio Networks: Networking, protocols and traffic performance*. West Sussex England: John Wiley, 2002.

APPENDIX

We show that given a set of packets, finding a packet bundling assignment with a minimal number is NP-complete.

Packet bundling assignment problem: Given a set of packets of varying sizes, a_1, a_2, \dots, a_n , can we assign them into k time slots such that the sum of the sizes of all packets assigned to the same time slot is less than or equal to the length L of the slot, where L is a constant?

To prove that it is NP-complete, we prove that the following Bin Packing Problem that is known to be NP-complete [8] can be reduced to our packet bundling problem in polynomial time.

Bin packing problem: Given a set of objects of variable sizes, O_1, O_2, \dots, O_n , where $0 < O_i \leq 1$, $1 \leq i \leq n$, can we pack them into b identical bins? We assume each bin has a capacity 1. We can assume the sizes O_1, O_2, \dots, O_n are rational numbers.

The polynomial reduction can be done in the following way. Let D be a denominator of O_1, O_2, \dots, O_n . We construct n packets with sizes $a_i = D \cdot O_i$, $0 < O_i \leq 1$. Obviously, a_i , $1 \leq i \leq n$ are integers. Let the length of a time slot $L = D$, and let the number of time slots $k = b$. It is obvious that the n objects can be packed into b bins if and only if the n packets can be assigned into k time slots. Therefore, the Packet bundling assignment problem is NP-complete.