QoS and Channel Aware Packet Bundling for VoIP Traffic in Cellular Networks

Jung Hwan Kim
Openet
Reston, VA 20190 USA
Email: junghwankim42@gmail.com

Baek-Young Choi and Cory Beard
Dept. Computer Science and Electrical Engineering
University of Missouri
Kansas City, Missouri 64110 USA
Email: {choiby,beardc}@umkc.edu

***Abstract*—We study the problem of multiple packet bundling to improve spectral efficiency in cellular networks. The packet size of real-time data, such as VoIP, is often very small. However, the use of time division multiplexing (TDM) limits the number of VoIP users supported, because a packet has to wait until it receives a time slot. Packet bundling can alleviate such a problem by sharing a time slot among multiple users. A recent revision of cdma2000 1xEV-DO introduces the concept of the multi-user packet (MUP) in the downlink to overcome limitations on the number of time slots. However, the efficacy of packet bundling is not well understood, particularly in the presence of time varying channels. We propose a novel QoS and channel aware packet bundling algorithm that uses adaptive modulation and coding. We show that channel utilization can be significantly increased by delaying some real-time packets slightly within their QoS requirements. We validate our study through OPNET simulation with a complete EV-DO implementation.**

I. Introduction

A growing demand for downlink-intensive applications such as Web browsing and file transfer over wireless networks, urges the need to use the wireless channel efficiently. Moreover, an emerging strong demand for delay sensitive data applications such as VoIP, wireless gaming and push-to-talk (PTT) over cellular networks, poses challenges on a network system to support a large numbers of simultaneous users while meeting their desired delay requirements.

Since the capacity of wireless systems is particularly constrained by the nature of location dependent and time varying channel conditions, careful attention needs to be paid to algorithms over wireless links in order to use the channel as efficiently as possible. In this work, we study the problem of multiple packet bundling to improve spectral efficiency in cellular networks. The packet size of real-time data, such as VoIP, is often very small. However, the use of time division multiplexing (TDM) limits the number of VoIP users supported, because a packet has to wait until it receives its time slot. The time slot cannot be made too small due to a relative MAC layer overhead for each time slot. Packet bundling can alleviate such a problem by sharing a time slot among multiple users.

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 wireless standards define the QoS framework and various types of service flows, but leave the QoS based packet scheduling and radio resource assignment undefined. Particularly, a 'multi-user packet' is among the improvements and expansions of EV-DO Rev A. That is, a downlink permits the access network to serve multiple users with the same physical and MAC layer packet. However, there is no guideline or recommended strategy in multiple packet bundling, and the efficacy of multi-user packets is not well understood, especially in the presence of location dependent and time varying channels. The concept of packet bundling is illustrated in Figure 1. Packets from multiple users or multiple packets from a single user may be combined together for a single time slot. Intuitively the bundling will increase channel utilization. Furthermore, it will decrease the average queueing delay of the VoIP packets, since later arriving VoIP packets do not have to wait for their time slot. An important aspect to consider, however, is bundling packets from MSs with different channel conditions. Advanced adaptive wireless systems employ channel measurement and feedback based rate control mechanisms such as the cdma2000 1xEV-DO system. In order for the bundled packet to be received reliably, the adaptive coding rate for the packet should correspond to the worst channel condition among the users. Then the channel utilization gain from packet bundling may deteriorate due to the lowest coding rate. One way to tackle the issue is to combine packets with the same or similar channel condition. A problem we observe from this approach is that at the time of bundling if there are not enough packets with the same or similar channel condition, the gain in the channel utilization may be marginal.

Our contributions are as follows. We first show that the optimal packet bundling algorithm that either maximizes channel utilization or minimizes queueing delay is a NP-complete problem. Then we propose a novel QoS and Channel aware packet Bundling (QCB) algorithm to jointly optimize QoS requirements and channel utilization with a simple approximation. QCB defers the bundling decision a little within the QoS requirement. In the meantime, the time slot can be used for best effort traffic, significantly increasing channel utilization. We compare the QCB scheme with bundling algorithms of two individual objectives, namely QoS Aware packet Bundling (QAB) and Channel Aware packet Bundling (CAB) schemes. We show that QCB enables high throughput as well as low

delay, achieving an optimal trade-off of the two extremes. We validate our study through OPNET simulation with a complete EV-DO implementation.

The remainder of this paper is organized as follows. Section II discusses related work on scheduling algorithms for general wireless networks as well as for EV-DO environments. Section III presents the background on the physical and MAC layer of the cdma2000 1x EV-DO Rev. A system relating to the issue of downlink packet bundling. Section IV discusses the hardness of a packet bundling problem and proposes approximation algorithms such as QCB, QAB, and CAB. The simulation setup and evaluation results are described in Section V. Section VI concludes the paper.

II. Related Work

A number of scheduling algorithms are available for wired networks including fair queueing, virtual clock, and earliest deadline first. However, these are not readily applicable to the wireless environment which has location dependent and time varying channel characteristics. Although there have been attempts to incorporate channel dependent features into schedulers for wired networks [5], they cannot effectively exploit the time-varying multiuser diversity gain. 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. This enables physical layer Adaptive Modulation and Coding (AMC) (see [2] for example). 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 [11] shows that Delay-Margin-based Scheduling nested with User-Channel-based Scheduling performs well both in delay and utilization metrics.

There are several studies on EV-DO downlink scheduling. Simulation studies on EV-DO VoIP capacity are presented in [19], [4]. [14] shows the trade-off between system throughput and delay with opportunistic scheduling with analysis and simulation of the EV-DO system. The authors in [6] 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 (PFsoft) shows the best performance over MAX rate algorithms. A forward link scheduling algorithm supporting the MUP scheme is proposed in [18]. This algorithm first selects the user's packet whose priority is the highest according to the PF algorithm. Then only packets with the same channel quality become the candidates for bundling with a higher priority given to VoIP packets. Otherwise, a single user packet (SUP) will be sent. The bundling ratio is limited by the available packets with the same channel condition. Our work differs from the above in that our scheduling algorithm jointly considers QoS and channel quality for packet bundling, and packets to MSs of different channel conditions may be bundled using AMC.

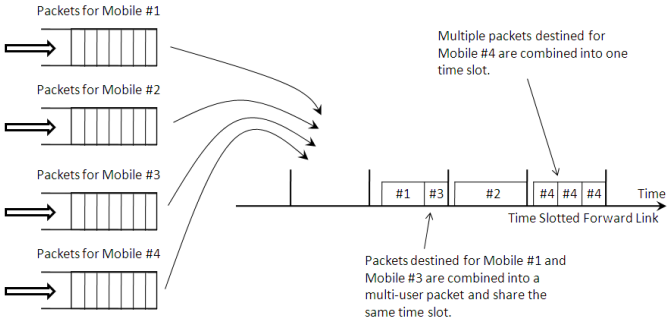

Fig. 1. The concept of packet bundling.

III. Background

In this section, we give an overview of the physical and MAC layers of the cdma2000 1xEV-DO Rev. A system relating to the issue of downlink packet bundling.

In a wireless system, signal strength is location dependent and time varying. It is subject to slow fading, fast fading and interference from other signals, resulting in degradation of the Signal to Interference-plus-Noise Ratio (SINR) [17]. A high SINR yields a high data rate and low error. A good SINR in cellular systems is achieved by using optimum rate and power control mechanisms.

In EV-DO networks, both CDMA and TDM are used in the downlink and CDMA is used in the uplink [9], [10]. The downlink channel is a single broadband link shared by all users in a cell. One user is allowed to receive data in a single time slot. The base station (BS) estimates each user's channel condition based on the feedback from individual mobile station(MS)'s measurements. The channel quality indication (CQI) feedback from the MS and the corresponding Adaptive Modulation and Coding (AMC) schemes are adopted in many current and future wireless standards including cdma 2000 1xEV-DO, WCDMA downlink, IEEE 802.16 broadband wireless access (WiMax), and the IEEE 802.11 Wireless LAN. In time slotted systems, the number of users supported is constrained theoretically. The maximum supported users or flows are limited by the number of time slots per second and packet arrival rates (Eq. (1)).

$$max\_supported\_users = \frac{no\_time\_slots/sec}{packet\_arrival\_rate/user} \quad (1)$$

For example, EV-DO revision A uses 1.25 MHz bandwidth with direct sequence spread spectrum (DSSS). The chip rate is 1.2288 Mchips/second, and the basic timing unit is 2048 chips. The channel slot time is 1.667 ms and there are 600 slots per second. Thus, it can serve a maximum of 600 packets per second. Suppose a voice coder generates a VoIP packet every 20 msec (i.e., maximum 50 packets/sec) and its average activity ratio is about 50%. Then, the maximum VoIP users supported in the EV-DO system is only 24 $(= 600/(50 \times 0.5))$. Meanwhile, the channel may go underutilized since the VoIP packet sizes are generally small (refer to Table III), and not able to fill the entire time slot.

TABLE I
ADAPTIVE MODULATION AND CODING SCHEMES IN CDMA2000 1X EV-DO REV. A DOWNLINK

DRC	Data rate (kbps)	Bits	Code Rate	Modulation
1	38.4	1024	1/4	QPSK
2	76.8	1024	1/4	QPSK
3	153.6	1024	1/4	QPSK
4	307.2	1024	1/4	QPSK
5	307.2	2048	1/4	QPSK
6	614.4	1024	1/4	QPSK
7	614.4	2048	1/4	QPSK
8	921.7	3072	3/8	8-PSK
9	1228.8	2048	1/2	QPSK
10	1228.8	4096	1/2	16-QAM
11	1843.2	3072	1/2	8-PSK
12	2457.8	4096	1/2	16-QAM
13	1586.0	5120	1/2	16-QAM
14	3072.0	5120	1/2	16-QAM

The CQI from the MS is called the Data Rate Control Channel (DRC) in the EV-DO system. The measured DRC value is fed back to the base station once every 1.667 msec using a reverse control channel. This slot size is short enough so that each user's channel quality stays approximately constant within one time slot, as it can be shown by computing the Doppler frequency of a mobile user at 2 GHz. In each time slot, one user is scheduled for transmission. Each user constantly reports to the base station its instantaneous channel capacity, i.e., the rate at which data can be transmitted if this user is scheduled for transmission.

Depending on the DRC feedback value, AMC schemes are adopted to support variable data rates for a more reliable transmissions for different mobile stations' channel environments. Modulation schemes are closely related to physical packet size. That is, if physical packet size is less than or equal to 2048, QPSK is used, if physical packet size is 3072, 8PSK is used, and if physical packet size is 4096 or 5120, 16QAM is used. Table I shows modulation and coding options in the EV-DO Rev. A downlink.

On the reverse link where multiple MSs send transmission concurrently, the EV-DO system capacity is limited by the interference level measured by RoT (Rise over Thermal). The RoT value is the total received power divided by the thermal noise value. The sector RoT value should be less than a threshold (7 dB is commonly used) most of the time to stabilize the system. The Base Station measures the sector RoT value and informs mobile stations with the RAB (Reverse Activity Bit) whether RoT is higher than the threshold or not, so that the uplink rate can be controlled.

Speech is encoded using a variable rate vocoder via the Enhanced Variable Rate Codec (EVRC) that generates VoIP traffic depending on speech activity. Since a frame duration is fixed at 20 ms, the number of bits per frame varies according to the traffic rate. 171 bits, 80 bits, 40 bits, and 16 bits are generated for full, half, 1/4, and 1/8 rate coding, respectively [1], [12]. The more detailed description can be found in cdma2000 specification [16].

The multi-user packet is a new feature of EV-DO Rev. A and it is designed to support more users per given time period. It is very important to support more users per given time period in real-time applications like VoIP because their delay deadlines can be better met with multi-user packets. VoIP application is the best fit to the multi-user packet, since the VoIP packets are generated frequently (every 20 ms) and their sizes are small. A bundled packet can be recognized by the preamble of the physical layer packet and the MAC header. Figure 2 shows single as well as bundled packet formats. Single user packet bundling of n packets (SUP-multiplex, (b)) has n bytes of header for the packet lengths of individual packets. With multi-user packet bundling (MUP-multiplex, (c)), 2 bytes are necessary for each packet to identify the MS within a sector and the packet length. All mobile stations decapsulate it as if a multicast packet, to extract the packet portion destined to itself. To be specific, it works as follows. In a single user packet (SUP-simplex or SUP-multiplex), a preamble of the physical layer packet is used to hold the MAC Index which represents the destination of the packet. In a multi-user packet (MUP), a preamble of the physical layer packet is used to hold modulation scheme and the MAC header is used to hold the MAC Index and packet size of individual packets(see Figure 2 (d)).

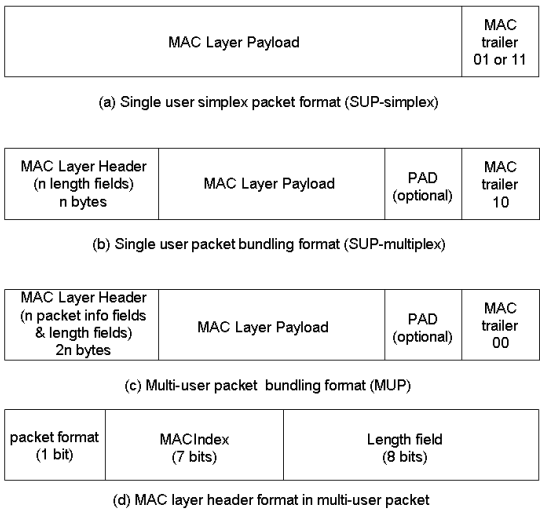

Fig. 2. Packet formats.

IV. MULTIPLE PACKET BUNDLING

In this section, we first show the hardness of the bundling problem. We then discuss QAB, CAB and QCB algorithms that approximately optimize QoS requirements, utilization, and both QoS and channel utilization, respectively. Efficient packet bundling may not always be beneficial due to diverse channel conditions observed at the mobile stations. When multiple packets are bundled, the modulation is used for a destination with the worst channel condition. A modulation for a worse channel means a lower bit rate to compensate for a higher error rate. Thus a multi-user packet may sacrifice data rates for all users based on the worst channel. Thus, bundling is not a trivial task and a careful decision needs to be made.

Remove the longest delayed VoIP packet from the queue and make a MUP;
while *the queue is not empty and the MUP is not full* **do**
  if *the coding of next longest delayed VoIP packet from the queue is not compatible with the MUP* **then**
    if *the MUP is too small to add the VoIP packet* **then**
      Exit the while loop;
    end
    else
      Change the MUP with the new coding;
      Remove the VoIP packet from the queue and add it to the MUP;
    end
  end
  else
    Remove the next longest delayed VoIP packet from the queue and add it to the MUP;
  end
end
if *the MUP is not full* **then**
  Add a BE packet to the MUP;
end

Algorithm 1: QoS Aware packet Bundling (QAB)

if *no VoIP packet* **then**
  Add a BE packet to a SUP;
end
else
  while *the queue is not empty and the MUP is not full* **do**
    Remove a VoIP packet from the queue and add it to the MUP with corresponding coding format;
  end
  foreach *defined MUP format* **do**
    if *the number of VoIP packets* $\geq B_{thresh}$ **then**
      Create a MUP using VoIP packets;
      if *the MUP is not full* **then**
        add a BE packet to the MUP;
      end
      Exit foreach loop;
    end
  end
  if *no MUP created* **then**
    add a BE packet to a SUP;
  end
end

Algorithm 2: Channel Aware packet Bundling (CAB)

A. Hardness of the problem

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

Packet bundling assignment problem: Given a set of packets of varying size, time slot, and an integer b, is there a bundling assignment or partition of the packets into time slots, with partition size less than b?

To prove that it is NP-complete, we prove that the following Bin Packing Problem that is known to be NP-complete [8], [13] can be reduced to our packet bundling problem in polynomial time.

Bin packing problem: Find a partition and assignment of a set of objects such that a constraint is satisfied or an objective function is minimized (or maximized). Specifically, determine how to put the most objects in the least number of fixed space bins. More formally, given a bin size V and a list $a_1, \ldots, a_n$ of sizes of the items to pack, find an integer B and a B-Partition of a set $S_1 \cup \ldots \cup S_B$ such that $\sum_{i \in S_k} a_i \leq V$ for all $k = 1, \ldots, B$.

The reduction is trivial in that the object and bin sizes correspond to the packet size and time slot interval, respectively, and the partition relates to the packet bundle assignment. Notice that this problem is easier than the problem of finding an optimal or minimal packet partition. If a minimal partition is known, simply computing its size and comparing it to B allows us to answer the question.

B. QoS Aware Packet Bundling (QAB)

With a QoS aware scheduling, a user whose packet is delayed the longest p_u^{*d} will be selected for service as below, when packet bundling is not used.[1][2]

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

where $d(p_u)$ is the delay of a packet of user u. The above equation can be extended to a set of packets for bundling, B^d as follows:

$$B^{*d} = \arg\max_{B^d} |B^d| \sum_{u \in B^d} d(p_u) \tag{3}$$

such that $\sum_{u \in B^d} L(p_u)/AMC(u) \leq T$. Where T is the time slot size, and $L(p_u)$ and $AMC(u)$ are the packet size and the AMC rate of user u, respectively.

As discussed earlier, finding such a set of packets for bundling is an NP-complete. Thus, we use an approximation algorithm called QAB as shown in Algorithm 1. The input is a queue of VoIP packets and the output is a packet bundling assignment. The QAB algorithm is similar to Earliest Deadline First (EDF) algorithm. Both algorithms are designed to serve real-time applications like VoIP. When there is not a real-time packet to bundle, a BE packet will be sent along. The packet size of BE traffic is often big enough for an entire time slot. For handling BE traffic, we use the PF algorithm for fairness as follows:

$$p_u = \arg\max_u CQI(u)\mu(u) \tag{4}$$

where $CQI(u)$ is the channel rate of user u determined by channel quality indicator, and $\mu(u)$ is a long term rate of user u.

[1] A packet will be dropped if the delay is greater than the requirement.

[2] We discuss QoS mainly in the context of delay parameter. However, it can be easily applied to other QoS parameters.

C. Channel Aware Packet Bundling (CAB)

As the channel condition varies depending on the time and the location of a user, the transmission data rate that a BS can send to an MS changes depending on the channel condition. Opportunistic scheduling that maximizes the channel utilization is to choose a packet p_u^{*c} whose channel rate $CQI(u)$ is the maximum. That is

$$p_u^{*c} = \arg\max_u CQI(u) \tag{5}$$

A natural extension of the scheme to packet bundling is to choose the set of packets , B^c that gives the maximum sum of CQIs within the time slot.

$$B^{*c} = \arg\max_{B^c} \sum_{u \in B^c} CQI(u) \tag{6}$$

subject to $\sum_{u \in B^c} L(p_u)/AMC(u) \leq T$. Since an algorithm that finds such a set of packets is NP-complete, a heuristic algorithm can be used to approximate the maximum rate bundling. A sketch of the CAB algorithm is shown in Algorithm 2. In order to better utilize the channel, *packets from the same or similar channel conditions* are bundled together. Since the worst AMC rate of the bundled packets will be the same or similar to the users' channel condition, the bundling ratio is high, resulting in efficient channel utilization. Also, for efficient handling of small size, real-time packets, it defines a bundling threshold, B_{thresh}, which is the minimum real-time data size or the number of packets that should be bundled. By limiting B_{thresh} as small, packets can be scheduled without being deferred, particularly, when there are little real-time packets to be bundled. Big B_{thresh} forces the real-time packets to be bundled with high bundling ratio, in order to better utilize the channel with BE traffic. Note that since the objective is only to maximize the utilization, it is impossible to 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.

while *delay of a VoIP packet* $\geq D_{thresh}$ **do**
  run QAB algorithm;
end
run CAB algorithm;

Algorithm 3: QoS and Channel aware packet Bundling (QCB)

D. QoS and Channel Aware Packet Bundling (QCB)

The main objectives of the QCB scheme are first to satisfy delay requirements of real-time packets, and then to utilize the wireless channel efficiently. We first define a maximum allowed delay, D_{thresh} that scheduling of real-time packets can be *deferred* in the queue without sacrificing 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. When the packets' delays are less than D_{thresh}, they attempt to utilize the channel efficiently by gathering packets of similar channel conditions that can be bundled together. The deferred scheduling of real-time packets makes room for opportunistic scheduling. For our experiments in Section V, we set D_{thresh} to be 25 ms, and B_{thresh} to be 4. We have varied the parameters and found that those values provide a good tradeoff between QAB and CAB. When B_{thresh} is 1, the QCB algorithm is the same as QAB. When D_{thresh} is 0, the QCB algorithm is reduced to the CAB algorithm. The pseudo-codes of the QCB algorithm are illustrated in Algorithm 3.

V. Evaluation

We have implemented the complete cdma2000 1xEV-DO system recommended by the 3GPP2 evaluation methodology [15] using OPNET. To the best of our knowledge, it is the first simulation that includes both a downlink and a uplink of the EV-DO system.[3] Even though our work is focused on downlink resource allocation and scheduling, the performance of the downlink is tightly coupled with uplink feedback and control mechanisms. Therefore, our implementation provides practical insights from the interplay of both links. In this section, we first describe the EV-DO simulation setup used in our study, and then discuss several prominent results of our extensive simulations.

A. Simulation Setup

As for cell interference, we implement a 19 cell wraparound model as depicted in Figure 3. It makes the interference environment more realistic than with a 7 cell model and is recommended in [15] as it considers second level interferences. With the wraparound model, the interference affects every cell in the simulation. Thus, we can collect statistical results from all cells rather than only from the center cell. In Figure 3, 19 white cells are our modeled cells. The other gray cells are imaginary cells that show how wraparound models work. For example, when we calculate interference of white cell 11, cell 10, 3, 12, 18, 19, and 15 give first level interferences and cell 16, 9, 2, 1, 4, 13, 17, 6, 7, 8, 14, and 5 give second level interferences. Each cell has three sectors. We evaluate the performance of all the 57 sectors to provide statistical results.

The path distance and angle used to compute the path loss and antenna gain of an MS at (x, y) to a BS at (a, b) are computed with Equations 7 and 8, respectively from [15].

$$Path\_loss = 28.6 + 35log10(d)dB \tag{7}$$

where d is the distance between BS and MS in meters.

$$A(\theta) = -min(12 * (\theta/70.0)^2, 20)dB \tag{8}$$

where $-180 \leq \theta \leq 180$.

[3]Previous EV-DO evaluation studies [3], [19] have been conducted on each link separately.

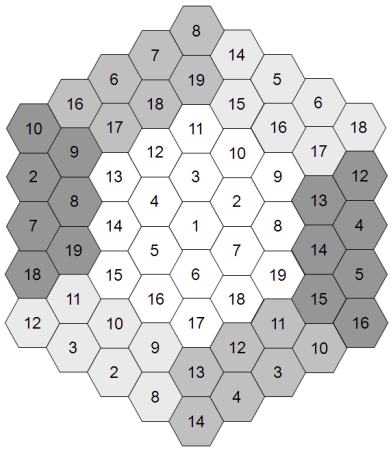

Fig. 3. 19 cell wraparound model. 19 white cells are our modeled cells. The other gray cells are imaginary cells that give interferences.

TABLE II
CHANNEL MODELS USED

Channel model	Multi-path model	No. of fingers (paths)	Speed (kmph)	Fading	Model assignment probability
Model A	Pedestrian A	1	3	Jakes	0.30
Model B	Pedestrian B	3	10	Jakes	0.30
Model C	Vehicular A	2	30	Jakes	0.20
Model D	Pedestrian A	1	120	Jakes	0.10
Model E (Station-ary)	Single path	1	0, f D=1.5 Hz	Rician Factor K = 10 dB	0.10

The distance used in the path loss between MS at (x, y) to BS at (a, b) is the minimum of the following.

$$\begin{aligned} min \quad \{ \quad & Dist\{(x,y),(a,b)\} \\ & Dist\{(x,y),(a+3R,b+8\sqrt{3}R/2)\} \\ & Dist\{(x,y),(a-3R,b-8\sqrt{3}R/2)\} \\ & Dist\{(x,y),(a+4.5R,b-7\sqrt{3}R/2)\} \\ & Dist\{(x,y),(a-4.5R,b+7\sqrt{3}R/2)\} \\ & Dist\{(x,y),(a+7.5R,b+\sqrt{3}R/2)\} \\ & Dist\{(x,y),(a-7.5R,b-\sqrt{3}R/2)\}\} \end{aligned}$$

where R is the radius of a circle that connects the six vertices of the hexagon.

We used five channel models as recommended in [15]. Channel models are randomly assigned to each mobile station. The probabilities that MSes take model A, B, C, D, and E are 0.3, 0.3, 0.2, 0.1 and 0.1, respectively. Table II summarizes the used channel models.

As the effectiveness of the scheduling algorithms would depend on the traffic mix, we evaluate the algorithms under various scenarios. We vary the number of VoIP sessions from 10 to 30 users. Additionally, 10 Best Effort (BE) sessions are added to observe the interplay of VoIP and BE traffic. For VoIP traffic, EVRC is used as mentioned in Section III. We also use silence suppression for VoIP packets, where a 1/8 rate packet is generated every 240ms in a silence mode. Robust Header Compression (RoHC) [7] is used as recommended in [16]. RoHC reduces an IP header from 40 types to just 3 bytes that leads to significant bandwidth savings. For BE traffic, FTP file downloads are performed for large files, so that the channels would never go idle. The uplink activity includes reverse activities of applications such as reverse direction VoIP (two way conversation) and TCP acknowledgements. Table III summarizes other simulation parameters.

TABLE III
SUMMARY OF PARAMETERS USED FOR SIMULATION

Parameter	Value
# of VoIP users/sector	10, 20, 30
# of BE users/sector	10
Bandwidth	1.25 MHz
Cell radius	1 Km
Maximum BS transmission power	20W (43 dBm)
Slot length	1.667 ms
VoIP packet length	5B $\sim$ 23 B after RoHC
Interval of VoIP packet generation	20 ms
Path loss exponent	3.5

B. Simulation Results

We first discuss the delay and throughput performances of bundling algorithms. Figure 4 compares average delays of VoIP traffic for QAB, CAB, and QCB schemes. The BE traffic throughput of the three schemes are shown in Figure 5. First, the BE throughput decreases as the number of VoIP users increases in all cases, because VoIP traffic receives priority over BE traffic. QAB performs the best in VoIP delay as it schedules based on the remaining time to meet the QoS. Meanwhile, CAB exhibits the most throughput in BE, maximizing channel utilization. Notice that despite the extra delay due to the deferred bundling time in QCB (maximum 25 ms), the performance is a lot closer to QAB than CAB. Meanwhile, in terms of throughput, our scheme shows a high performance close to CAB due to bundling efficiency. Figures 4 and 5 show a good performance tradeoff between delay and throughput of the QCB algorithm. In fact, if we can allow even more VoIP delay depending on remaining time to deadline, we can get more BE throughput via exploiting better channel diversity. However, the trade-off between delay and throughput is achieved optimally with around 25 ms bundling delay, for the given parameters of the traffic load. Due to space limitation, we do not show the results.

Figure 6 shows the interesting behavior of delay cumulative distribution functions (CDFs) of VoIP packets for SUP multiplex and MUP schemes. In SUP multiplex, VoIP delay increases when the number of users increases. Meanwhile VoIP delay decreases with MUP, as the number of users increases. This is because in SUP multiplex, each user takes turns in the use of time slots and the period becomes longer with the increased number of users. On the other hand, in MUP, the more VoIP packets from the increased number of users makes the bundling easier with little need to wait, thus enhancing the multi-user diversity gain. In both QCB-MUP and QCB-SUP multiplex cases, the CDFs show a longer tail

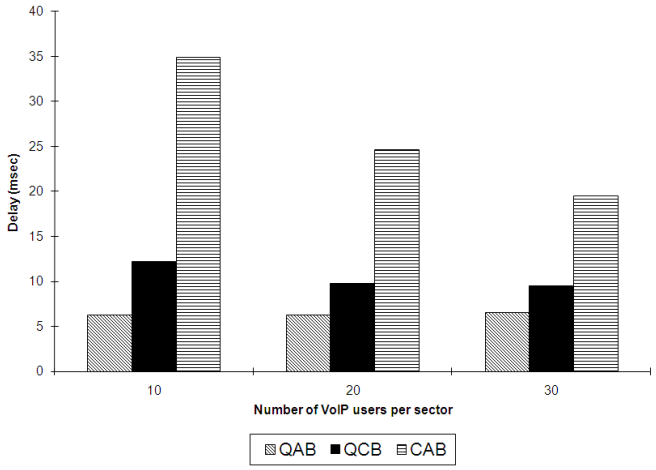

Fig. 4. Comparisons of bundling algorithms for average VoIP traffic delay

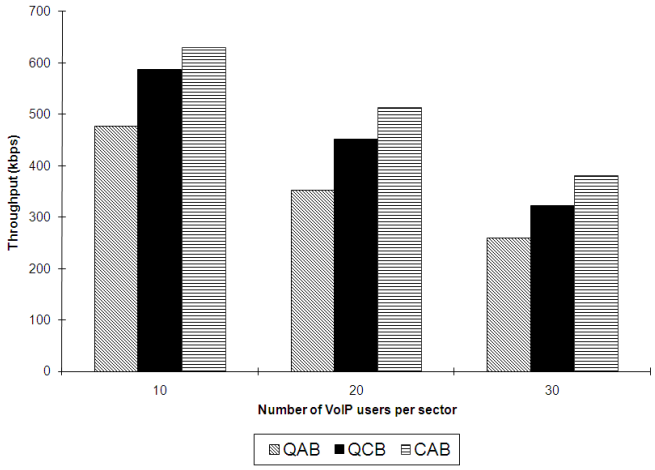

Fig. 5. Comparisons of bundling algorithms for BE throughput

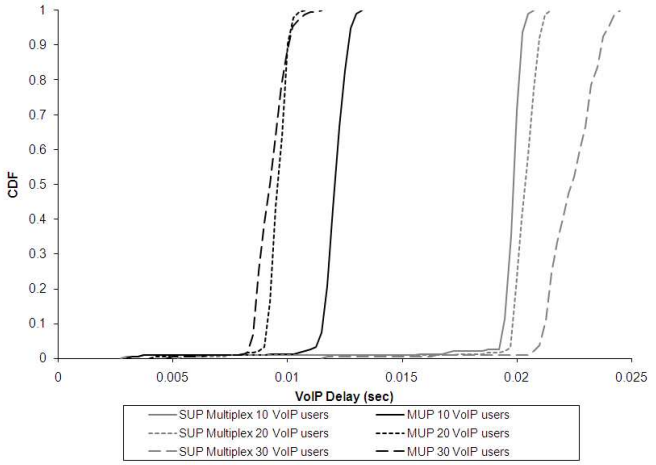

Fig. 6. Empirical Cumulative Density Functions of VoIP packet delays for SUP multiplex and MUP(variants of QCB)

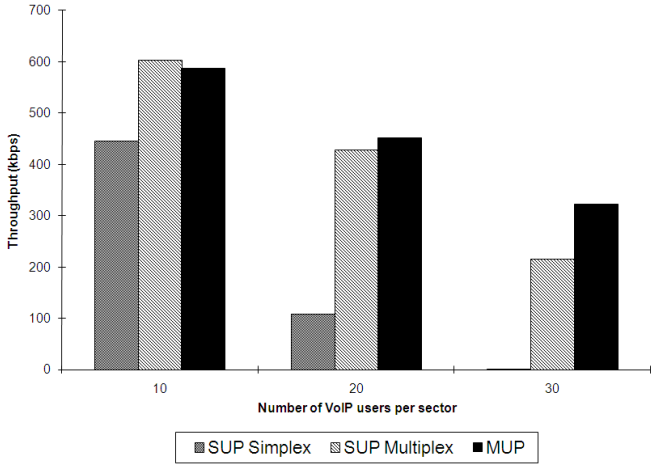

Fig. 7. Throughput of BE for SUP-simplex (no-bundling), SUP multiplex and MUP (variants of QCB)

for a greater number of VoIP users. VoIP delay generally decreases when number of VoIP users increases because the chance of bundling is higher. However, some VoIP packets have a higher delay when the number of VoIP user increases because the chance of congestion (many VoIP packets existing in the queue) is higher.

Figure 7 compares the BE throughput of the SUP simplex, the SUP multiplex, and the MUP. First, the BE throughput decreases as the number of VoIP users increases, since the higher priority is given to VoIP over the BE traffic. The decrease of BE throughput is more prominent in the SUP simplex than in bundling schemes. The throughput of packet bundling using either the SUP multiplex or the MUP degrades gradually as they attempt to maximize channel utilization with higher rates of bundling. Particularly, the SUP multiplex shows higher BE throughput with a small number of VoIP users, and the MUP wins over the SUP multiplex with a large number of VoIP users. It shows that the efficiency of the MUP increases when number of users grows, as it takes advantage of multi-user diversity better. As the MUP format is decided by the worst DRC values of MSs whose packets are bundled, it is more likely to find similar DRC users as the number of VoIP users increases.

TABLE IV
AVERAGE PACKET LOSS (%)(PACKET ERROR + DROP)

No. VoIP users/sector	SUP Simplex	SUP Multiplex	MUP
10	0.21	0.37	0.25
20	1.70	0.28	0.23
30	25.56	0.21	0.13

Table IV gives the average packet drop rates. It clearly shows that the the SUP simplex cannot handle much VoIP traffic (30 users) with very high packet loss rates over 25 %. This figure shows we need bundling if we want to handle VoIP traffic. The SUP multiplex and the MUP both are fine in drop probability.

Now let us consider the overhead of extra packet headers incurred by our proposed packet bundling, QCB. When single

user packet bundling is used (See Figure 2, (b)) for n packets, the excess header size is $8 \times n$ bits. With multi-user packet bundling, it is $16 \times n$ bits. With our simulation using 30 VoIP users and 10 BE users per sector and 25 ms delay max allowance, the average number of bundled packets in a SUP packet was 1.9, and the average size of the bundled VoIP packets was 486 bits. Meanwhile, the average number of bundled packets in a MUP packet was 4.3, and the average size of the bundled VoIP packets was 1429 bits. Therefore, the overheads of SUP and MUP are $1.9 \times 8/486 = 3.1\%$ and $4.3 \times 16/1429 = 4.8\%$ respectively which is negligible compared to the huge utility gain.

VI. Conclusions

We have proposed a joint QoS and Channel Aware Packet Bundling (QCB) algorithm for VoIP packets to improve spectral efficiency in cellular networks. The packet size of real-time data such as VoIP is often very small, leaving channels underutilized in TDM cellular systems. Packet bundling could improve the channel utilization in such networks. However, a careful treatment should be paid due to location dependent and time varying channel characteristics of wireless networks. Since the packet bundling algorithm is a NP-complete problem, we introduce approximation algorithms, namely QoS Aware Packet Bundling (QAB), Channel Aware Packet Bundling (CAB) and QCB. We have validated the efficacy of the approximation algorithms, through extensive simulations of a complete EV-DO implementation to our knowledge. We have shown that the QCB scheme out-performs QAB and CAB, thus truly maximizes a multi-user/traffic diversity gain, as it achieves a high throughput for BE traffic while keeping a low delay. We have further investigated the behavior of QCB variants, and found that the QCB-Multi-User-Packet (QCB-MUP) is more effective when there are larger numbers of VoIP users and the QCB-Single-User-Packet-multiplex (QCB-SUP-multiplex) demonstrates more BE-throughput and a lower overhead with small numbers of VoIP users.

As for future work, we plan to investigate the performance of QCB when multiple flows per node are allowed. With multiple flows per node, we expect the BE throughput of the QCB-SUP multiplex will be improved more than the current results show. With our current work, when VoIP packets are sent in the SUP multiplex case, only VoIP packets are sent because the node doesn't have any BE traffic. When multiple flows are permitted, VoIP and BE traffic may be sent together leading to a better channel utilization in the QCB-SUP multiplex. Multiple flows however, are not expected to make differences in the performance of the QCB-MUP scheme. We are also working on extending the current work to multi-carrier wireless environments.

Acknowledgment

The authors would like to thank John Kim and Shiva Narayanabhatla at Sprint-Nextel for their practical insights and information for the implementation.

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] N. Bhushan, C. Lott, P. Black, R. Attar, Y.-C. Jou, M. Fan, D. Ghosh, and Jean Au. 1xEV-DO Revision A: Physical and MAC Layer Overview. *IEEE Communications Magazine*, 44(2):75–87, Feb. 2006.

[4] 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 Commmunications*, 24(1):36–45, 2006.

[5] Y. Cao and V. Li. Scheduling algorithms in broadband wireless networks. *Proc. IEEE*, 89(1):76–87, Jan 2001.

[6] 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.

[7] M. Degermark, B. Nordgren, and S. Pink. IP Header Compression (IPHC). RFC2507.

[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] 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.

[12] TIA IS-127. Enhance Variable Rate Codec (EVRC) 8.5 kbps Speech Coder.

[13] David S. Johnson, Alan J. Demers, Jeffrey D. Ullman, M. R. Garey, and Ronald L. Graham. Worst-Case Performance Bounds for Simple One-Dimensional Packing Algorithms. *SIAM Journal on Computing*, 3(4):299–325, 1974.

[14] Roshni Srinivasan. *Scheduling in Packet Switched Cellular Wireless Systems*. PhD thesis, University of Maryland, College Park, 2004.

[15] 3GPP2 C.R1002-0 v1.0. cdma2000 Evaluation Methodology . http://www.3gpp2.org/public_html/specs/C.R1002-0_v1.0_041221.pdf, Dec. 2004.

[16] 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.

[17] B. H. Walke. *Mobile Radio Networks: Networking, protocols and traffic performance*. West Sussex England: John Wiley, 2002.

[18] Qu Yajiang, Wang Chunye, and Wang Xiaoyi. Scheduling for multi-user packet in CDMA2000 1x EV-DO. In *Proc. IEEE International Conference on Mobile Technology, Applications and Systems*, Nov. 2005.

[19] M. Yavuz, S. Diaz, R. Kapoor, M. Grob, P. Black, Y. Tokgoz, and C. Lott. VoIP over cdma2000 1xEV-DO revision A. *IEEE Communications Magazine*, 44(2):50–57, Feb. 2006.