Quality-of-Service Provisioning and Efficient Resource Utilization in CDMA Cellular Communications

Hai Jiang, *Student Member, IEEE*, Weihua Zhuang, *Senior Member, IEEE*,
Xuemin (Sherman) Shen, *Senior Member, IEEE*, and Qi Bi, *Senior Member, IEEE*

***Abstract*—One of the major challenges in supporting multimedia services over Internet protocol (IP)-based code-division multiple-access (CDMA) wireless networks is the quality-of-service (QoS) provisioning with efficient resource utilization. Compared with the circuit-switched voice service in the second-generation CDMA systems (i.e., IS-95), heterogeneous multimedia applications in future IP-based CDMA networks require more complex QoS provisioning and more sophisticated management of the scarce radio resources. This paper provides an overview of the CDMA-related QoS provisioning techniques in the avenues of packet scheduling, power allocation, and network coordination, summarizes state-of-the-art research results, and identifies further research issues.**

***Index Terms*—Code-division multiple-access (CDMA), intercell coordination, multimedia services, packet scheduling, power allocation, quality-of-service (QoS), soft handoff.**

I. Introduction

THE PAST decade has witnessed the success of code-division multiple-access (CDMA) technology in the second-generation cellular systems, due to its promising advantages such as universal frequency reuse, soft handoff, inherent diversity, soft capacity, and high spectrum efficiency. CDMA is also the major multiple access technology for the third-generation wireless communication systems and beyond. On the other hand, it is widely accepted that the future wireless access networks are expected to converge into an all-Internet protocol (IP) architecture. Fig. 1 shows an all-Internet protocol (IP) architecture including CDMA cellular networks, 802.11 wireless local area networks (WLANs), and Bluetooth/ultra-wideband (UWB) personal area networks (PANs). The base station (BS) or the access point (AP) provides the mobile stations (MSs) with wireless Internet access. With the rapid growth of the Internet, the demand for fast and location-independent mobile multimedia services will be steadily increasing in future IP-based CDMA cellular systems. Typical applications include voice-over-IP (VoIP), videoconferencing, video streaming, distance learning, web browsing, file transfer, and e-mails. Therefore, multimedia communications over IP-based CDMA cellular networks have recently received significant interests from both industry and academia.

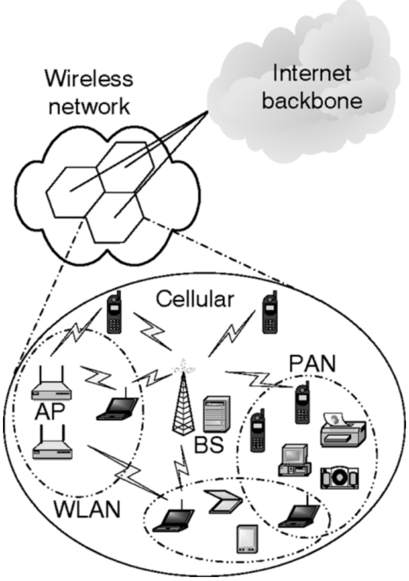


Fig. 1. The future all-IP heterogeneous wireless networks.

One major challenge in multimedia services over CDMA cellular networks is the quality-of-service (QoS) provisioning with efficient resource utilization. Compared with the circuit-switched voice service in the second-generation CDMA systems (i.e., IS-95), heterogeneous multimedia applications in future IP-based CDMA networks require more complex QoS provisioning and more sophisticated management of the scarce radio resources. QoS can be classified according to its implementation in the networks, based on a hierarchy of three different levels: bit level, packet level, and call level. The transmission accuracy, transmission rate (i.e., throughput), timeliness (i.e., delay and jitter), and fairness are the main consideration in this classification. This classification also reflects the QoS categories from a technical viewpoint.

- Bit-level QoS—To ensure some degree of transmission accuracy, a maximum bit-error rate (BER) for each user is required. The BER guarantee can be realized by satisfying a required bit-level signal energy to interference-plus-noise density ratio, i.e., E_b/I_0. The one-to-one mapping of BER to E_b/I_0 depends on channel characteristics, modulation, channel coding, diversity, and receiver design.
- Packet-level QoS—As real-time applications (such as VoIP and videoconferencing) are delay-sensitive, each packet should be transmitted within a delay bound. On the other hand, data applications are usually delay-tolerant,

Manuscript received September 24, 2004; revised May 26, 2005.

H. Jiang, W. Zhuang, and X. Shen are with the Centre for Wireless Communications, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: hjiang@bbcr.uwaterloo.ca; wzhuang@bbcr.uwaterloo.ca; xshen@bbcr.uwaterloo.ca).

Q. Bi is with Bell Laboratories, Lucent Technologies, Whippany, NJ 07981 USA (e-mail: qbi@lucent.com).

Digital Object Identifier 10.1109/JSAC.2005.858897

and throughput is a better QoS criterion. Each traffic type can also have a packet loss rate (PLR) requirement.

- Call-level QoS—In a cellular system, a new (or a handoff) call will be blocked (or dropped) if no sufficient capacity. From the user's point of view, the handoff call dropping is more disturbing than new call blocking. Effective call admission control (CAC) is necessary to guarantee a blocking probability bound and a smaller dropping probability bound.

To realize QoS provisioning in a CDMA cellular environment, there exist many challenges: 1) limited radio resources; 2) the time-varying wireless channel; 3) limited battery power supply of the MSs; 4) interference-limited capacity (leading to complex resource allocation); 5) possible service disruption and even call dropping during handoff.

To guarantee the bit-level and packet-level QoS requirements of the MSs, an effective packet scheduling with appropriate power allocation is needed. Specifically, the power levels of all the MSs are managed in such a way that each MS achieves the required E_b/I_0, and the transmissions from/to all the MSs are controlled by a scheduler to meet the delay, jitter, throughput, and PLR requirements. The order of packet transmissions for multimedia traffic has a great impact on the efficiency and performance of a CDMA cellular system. The design of a packet scheduler involves balancing a number of conflicting objectives. For different types of multimedia traffic, different scheduling policies can be applied, focusing on the corresponding main QoS criteria of the traffic types.

Call-level QoS can be guaranteed by an effective CAC scheme. CAC is also critical for packet scheduling (with power allocation) to provide bit-level and packet-level QoS. Interested readers may refer to [22], [48], [69], [71], [75] and references therein.

The capacity of a CDMA system is interference-limited. In a cellular environment, each MS will experience intracell and intercell interference. While intracell interference can be managed effectively by packet scheduling and power allocation, the intercell interference can be controlled in a systematic manner by means of coordination among the cells. Further, soft handoff is an inherent property of cellular CDMA, and requires coordination among neighboring cells in resource allocation to the MSs.

Techniques for multimedia services over CDMA cellular networks have been developed along many avenues. This paper focuses on QoS provisioning and resource utilization issues, summarizes state-of-the-art research results, and provides insights to further research work. The remaining of this paper is organized as follows. Section II describes the packet scheduling strategies with power allocation. In Section III, network coordination mechanisms are presented. Further discussion on differentiated services and cross-layer design is given in Section IV, as well as the conclusion remarks.

II. Packet Scheduling With Power Allocation

In a CDMA system, radio spectrum of the wireless channel is shared by all the MSs, which introduces interference. For each MS, two kinds of interference exist: intracell interference caused by simultaneous transmissions from/to other MSs in the same cell, and intercell interference originating from signals in other cells. A packet scheduling scheme is essential to schedule packet transmission and control interference in order to meet various QoS requirements such as delay bound, throughput, and PLR bound. A centralized scheduler at the BS benefits from more processing power and available information than a distributed one. For the downlink, the BS has the information of traffic status of each MS. For the uplink, each MS needs to send a transmission request upon new packet arrivals and update its link status to the BS. The request and update information is stored in its corresponding profile at the BS, as shown in Fig. 2. The request and update messages can be transmitted in a request channel, or piggybacked in the transmitted uplink packets to avoid possible contention in the request channel. The BS responds by broadcasting transmission decisions to MSs [32], [71], [72].

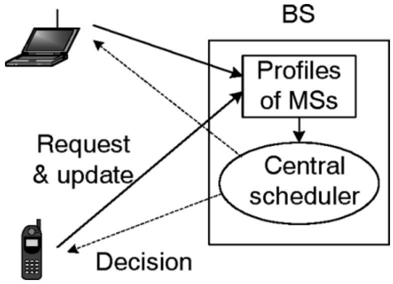


Fig. 2. Centralized scheduler for the uplink transmission.

To control the interference and overcome the well-known near–far problem, an appropriate power level should be allocated to each MS to guarantee a certain level of transmission accuracy at the bit level [79]. Consider the uplink with nonorthogonal simultaneous transmissions from different MSs. Let N denote the number of active MSs in the cell. MS i $(i = 1, \ldots, N)$ is characterized by E_b/I_0 requirement Γ_i, bit transmission rate R_i, and channel gain h_i. The transmit power vector, $\mathbf{P} = (P_1, \ldots, P_N)$, can be obtained by solving the following linear inequalities [1]

$$\frac{W}{R_i} \frac{h_i P_i}{\sum_{j \neq i} h_j P_j + I^o + \nu} \geq \Gamma_i, \quad i = 1, \ldots, N \tag{1}$$

with constraints

$$P_i \leq P_i^{\max} \tag{2}$$

where W is total uplink bandwidth used by all the MSs, I^o and ν are intercell interference and background noise power, respectively, and $P_i^{\max}$ is the transmit power limit of MS i. After some algebra, the following constraint can ensure there exists a feasible transmit power vector [3], [62]:

$$\sum_{i=1}^{N} \left(1 + \frac{W}{\Gamma_i R_i}\right)^{-1} \leq 1 - \delta_u \tag{3}$$

[1] For uplink with interference cancellation techniques, as the perceived intracell interference may be alleviated or eliminated, the inequalities (1) should be changed accordingly. For example, in the successive interference cancellation technique, the decoding order of MSs at the BS affects the achieved E_b/I_0 of the MSs in inequalities (1), as interference from previously decoded MSs is successively alleviated or eliminated in the composite received signal at the BS, i.e., the later decoded MS signals are immune to the previously decoded ones [44].

where

$$\delta_u = \frac{I^o + \nu}{\min_{1\le i\le N}\left[h_i P_i^{\max}\left(\frac{W}{\Gamma_i R_i}+1\right)\right]}. \tag{4}$$

Constraint (3) can also be considered as the instantaneous capacity region for the given link conditions. The term

$$\beta_i = \left(1 + \frac{W}{\Gamma_i R_i}\right)^{-1} \tag{5}$$

can be viewed as the *power index* [3] or the *effective bandwidth* [10] of MS i, while $1-\delta_u$ is the *total power index*, or *total effective bandwidth.*

The downlink capacity model is quite different from that of the uplink, as orthogonal codes are usually used for transmissions to all the MSs within a cell. However, the orthogonality among the signals even with synchronous transmissions in the downlink cannot be kept perfectly taking into account the transmission delay spread in a multipath propagation environment. This can be addressed by associating each MS with an orthogonality factor $\theta_i \in [0,1]$, where $\theta_i = 1$ means perfect orthogonality observed at MS i. The values of the orthogonality factor depend on the location of the MSs and the multipath environment, with a typical value in [0.4, 0.9] [29]. For presentation simplicity, to discuss the power allocation in the downlink, we use the same notations as those for the uplink. The transmit power vector $\mathbf{P} = (P_1,\ldots,P_N)$ can be obtained by solving the following linear inequalities:

$$\frac{W}{R_i}\frac{h_i P_i}{(1-\theta_i)h_i\sum_{j\ne i}P_j + I_i^o + \nu_i} \ge \Gamma_i, \quad i=1,\ldots,N \tag{6}$$

with constraint

$$\sum_{i=1}^{N} P_i \le P_d^{\max} \tag{7}$$

where $P_d^{\max}$ is the maximum transmit power at the BS. The condition for a feasible power vector $\mathbf{P} = (P_1,\ldots,P_N)$ allocation is [9]

$$\sum_{i=1}^{N}\frac{1}{1+\frac{W}{(1-\theta_i)R_i\Gamma_i}}\left(1+\frac{I_i^o+\nu_i}{(1-\theta_i)h_i P_d^{\max}}\right) \le 1. \tag{8}$$

To guarantee the E_b/I_0 requirement for each and every MS, inequality (3) or (8) should hold at all times.

For CDMA systems, the scheduling strategies can be classified into two groups: pure code scheduling and time scheduling. The pure code scheduling is for circuit switching, where an admitted MS can transmit continuously if it has traffic to send. The advantage of the pure code scheduling scheme is its implementation simplicity with modest computation complexity and signaling overhead. The voice or video traffic under pure code scheduling experiences no access delay because of the continuous transmission. However, this also leads to loss of flexibility and efficiency. On the other hand, time scheduling is for packet switching, where the scheduler determines which MSs in service can transmit in each time slot and the associated power levels. By time scheduling, an MS can achieve high instantaneous bit rate but in short periods [29], [59]. In the following, various time scheduling schemes are discussed in detail.

A. Time Scheduling for Nonreal-Time Data Traffic

One main focus of the third-generation wireless systems is to provide spectrally efficient wireless data services [11]. The emerging all-IP wireless architecture should accommodate the demand for high-speed packet data services in wireless/mobile environments. The relative delay-tolerant nature of data applications provides the possibility of flexible and efficient use of the spectrum via time scheduling. Specifically, if a data MS transmits at a high rate, the instantaneous interference to other MSs is high due to the required high transmission power; on the other hand, if a low transmission rate is used by a data MS, it prolongs the time duration over which the MS generates interference to others. Therefore, to efficiently mitigate the effect of the interference, an effective scheduling scheme is needed to determine the delivery order, the transmission rates, and power levels of packets from/to data MSs in order to achieve the best performance, e.g., maximum capacity in terms of maximum total throughput and/or minimum total consumed power, under the QoS constraints.

It is expected that data traffic will be mainly located in the downlink with applications such as file downloading and web browsing. In the downlink, if total transmit power of the BS for all the MSs is P_d and MS i is assigned a fraction φ_i ($\sum_i \varphi_i = 1$) of it, and the received E_b/I_0 at MS i is the same as the required Γ_i, then from (6), the transmission rate of MS i is [9]

$$R_i = \frac{W}{\Gamma_i}\cdot\frac{\varphi_i P_d h_i}{(1-\theta_i)(1-\varphi_i)P_d h_i + I_i^o + \nu_i} \tag{9}$$

in a nonfading channel. Consider the *one-at-a-time* scheduling case, i.e., at any instant, only traffic to one MS is transmitted with full power P_d, and MS i gets access to the channel only for a fraction φ_i of the total time. Then, the average achieved transmission rate of MS i is given by [6]

$$R_i^* = \varphi_i\cdot\frac{W}{\Gamma_i}\cdot\frac{P_d h_i}{I_i^o+\nu_i}. \tag{10}$$

A comparison of (9) and (10) shows that [6]

$$\begin{cases} R_i^* > R_i, & \text{for } 0 \le \theta_i < 1 \\ R_i^* = R_i, & \text{for } \theta_i = 1 \end{cases}. \tag{11}$$

This implies that the one-at-a-time transmission can achieve the largest overall system throughput against simultaneous transmissions of the MSs, except in the perfect orthogonality case, where the two transmission policies are equivalent in terms of the system throughput. If the optimal point is defined as the minimum total energy consumption under average transmission rate constraints, the one-at-a-time scheduling also has the merit of minimizing the required transmission energy, thus reducing interference to other cells [9].

Similarly, in the uplink, there is an increasing demand for data applications such as data file uploading, ftp service, and short message service (SMS). Under the E_b/I_0 constraints, permitting only a limited number of delay-tolerant data MSs to transmit at a time and using the round-robin mode lead to significant per-MS throughput gains [61]. Consider the transmission rate of an MS as a function of received signal-to-interference-plus noise ratio (SINR) $\mathcal{R}(\mathrm{SINR})$, and use an equivalent definition of power index (5), $\beta = (\mathrm{SINR}/(1+\mathrm{SINR}))$.

The transmission rate can be viewed as a function of the assigned power index $\mathcal{R}(\beta)$, which is shown to be convex [45]. This means that the one-at-a-time scheduling is favored to improve per-MS throughput performance. However, if the allowable peak transmit power is very small for each MS, single transmission may not be able to fully utilize the total link capacity, and simultaneous transmissions are favored under this condition [45], [63].

In the optimal one-at-a-time scheduling policy, MS i is allowed to transmit at rate

$$R_i = \frac{W}{\Gamma_i} \cdot \mathrm{SINR}_i \tag{12}$$

(which is an equivalent form of (9), for both downlink [6], [9] and uplink [10], [62]) under the assumptions that: 1) the spreading gain is large enough for spectrum spreading; 2) different transmission rates are provided by adapting the processing gain [8]; and 3) any value of the variable processing gain is achievable. However, assumption 3) may not be always valid. In the case where only a set of discrete transmission rates is allowed, the maximum system capacity can be achieved by a certain level of simultaneous transmissions. This is because the maximum value in the discrete transmission rate set may not fully utilize the link capacity in the one-at-a-time scheduling case [35].

B. Time Scheduling With Delay Bound

Real-time traffic (i.e., voice and video) normally has a strict delay bound requirement. Packets with transmission delay exceeding the bound are considered useless and discarded. Traditionally data traffic is deemed delay-insensitive. However, this may not be always true in practice. For example, for web browsing, packets will be discarded if they cannot be delivered successfully within a deadline. Furthermore, in order to provide an end-to-end error-free service to data traffic, transmission control protocol (TCP) is usually employed at the transport layer. A large packet delay may push TCP to timeout. Therefore, TCP performance will degrade severely if the packet delay is not controlled. In this context, it is desired that data traffic also has a delay bound, which can be much larger than that of voice or video.

To provide a generic scheduling model to delay-bounded voice/video/data traffic, a wireless multimedia access control protocol with BER scheduling (termed WISPER) [1] has the ability to support a number of multimedia traffic classes with different BER requirements. WISPER is based on a slotted CDMA system, where an uplink frame consists of a request slot for MSs to issue transmission requests and a few packet slots, and a downlink frame includes a control slot for the BS to send transmission permissions and several packet slots. Traffic arrives in batches. Each active batch has a priority value proportional to the remaining packet number in the batch and inversely proportional to the remaining time before the batch expires. Packets are transmitted in the descending order of the priority values among all active batches. Packets with the same or similar BER requirements are transmitted with the same received power level and in the same time slot whenever it is possible. This provides significant performance advantages in the presence of power control imperfection [72]. However, the maximum number of packets accommodated in a slot is determined by the most stringent BER requirement of the allocated packets, thus resulting in underutilization of a time slot. To overcome this shortcoming, different power levels should be assigned in each slot based on the BER requirements of the allocated MSs. Specifically, minimum power levels for all MSs to transmit in a slot can be determined based on the equalities in (1). Under this power allocation strategy, resource underutilization in each slot is avoided and better system performance is achieved [71], at the cost of more complex power allocation.

In terms of delay performance, nonpreemptive earliest deadline first (EDF) is the delay-optimal scheduling policy for a single server [24]. Among all the eligible packets, the scheduler selects a packet with the earliest departure deadline and transmits this packet nonpreemptively. Therefore, it is plausible to apply nonpreemptive EDF to a CDMA system supporting multimedia traffic with delay bounds. The EDF policy has shown its effectiveness in many CDMA applications [66], [77]. The multiple code channels in CDMA networks can be viewed as multiple servers. If multiple MSs are scheduled to transmit simultaneously by the servers, different power levels need to be assigned so as to avoid resource underutilization. However, the scheme is sensitive to imperfect power control. A solution is to make information bits simultaneously transmitted over the parallel code channels in each slot always belong to the same packet, thus, with the same power level and being stable in the presence of power control imperfection. Specifically, the incoming traffic is selected to be served in the order of nonpreemptive EDF. Each selected packet is equally partitioned into a number of parts to be transmitted by multiple code channels, and the partitioning number is determined by the BER requirement of the packet's class [14]. This mechanism retains the delay-optimal property of nonpreemptive EDF. One drawback is that each transmitter needs sufficient instantaneous power supply to support multiple code channel transmissions.

For bursty traffic, EDF may cause unfairness among MSs as only the delay bound is considered. In EDF, the packet dropping occurs when all the delay bounds cannot be satisfied due to insufficient radio resources over the period. When this happens, it is essential to distribute packet dropping among all the MSs in an appropriate way, as different traffic types have different levels of packet loss tolerance. This means a packet scheduling algorithm with fair packet loss sharing (called FPLS) [32] is needed. The FPLS scheduler can be designed such that each MS experiences packet dropping fairly according to its PLR requirement. Therefore, when an MS achieves QoS satisfaction (in terms of delay bound, BER, and PLR), so do all the others. The main disadvantage of FPLS is its computational complexity.

To guarantee an absolute delay bound, the well-known generalized processor sharing (GPS) [57], [58] is effective if all the arrival traffic is regulated, and is discussed as follows.

C. Wireless Fair Scheduling

To share network resources among a number of traffic sessions (each originating from or being destined to an MS), GPS is an ideal fair scheduling discipline, originally proposed for wireline networks. The basic principle of GPS is to assign a fixed weight to each session, and allocate bandwidth for all the sessions according to their weights and traffic load. For each backlogged session, GPS can provide a minimum service rate and guarantee a tight delay bound if traffic of the session is regulated by a leaky bucket. Isolation from ill-behaved sessions can also be provided by GPS [57], [58].

The main drawback of the GPS principle comes from its two assumptions: infinitely divisible traffic in the fluid-flow traffic model and simultaneous multiple transmissions from multiple sessions. The ideal GPS principle is not realizable in practical systems, especially, in a time-division multiple-access (TDMA) system where no parallel transmissions are allowed. Some packet-based variants of GPS are proposed for wireline or wireless TDMA systems, such as packet-by-packet GPS [also known as weight fair queueing (WFQ)] [57], worst-case fair weighted fair queueing (WF^{2}Q) [7], and idealized wireless fair queueing (IWFQ) [52]. On the other hand, CDMA is inherent to support simultaneous multiple transmissions with variable transmission rates, which makes it more effective to follow the GPS principle than TDMA. To support multimedia traffic, the code-division GPS (CDGPS) scheme [73] can realize the general principle of GPS successfully and efficiently in CDMA cellular networks.

To implement GPS in CDMA cellular networks, the resource should be defined first. Two methods can be considered for the definition.

- Similar to the case where GPS is applied to TDMA systems, the total transmission rate can be used as the system resources [49], [73]. However, in a CDMA system, the total transmission rate varies, affected by the BER requirements and the link quality levels of the allocated sessions. Such a varying total allowable transmission rate can be referred to as *soft capacity* [73]. For the uplink scheduling, the soft capacity can be determined based on (3).
- From (3), since the power indices of all the allocated MSs at any instant are linearly bounded, it is convenient to incorporate power index into the resource definition. The total system resources can be defined as the total power index $(1-\delta_u)$ [70] (or total power index times the chip rate [4]). The CDMA-based GPS implementation based on these resource definitions is simple and intuitive. However, it is unlikely to lead to a proportional transmission rate.

From (4), it can be seen that the total power index is constrained by the MS with $\min\{h_i P_i^{\max}((W/(\Gamma_i R_i)) + 1)\}$. If an MS experiences a large link loss, the total power index can be reduced significantly, even below a target value. In this case, a larger spreading gain can be assigned to this MS (which results in a lower transmission rate), making the total power index equal to or larger than the target value. This will lead to a smaller power index of this MS (named a *lagging MS*), and larger power indices and transmission rates of other MSs (named *leading MSs*) in a good channel condition. This unfairness can be compensated as the leading MSs will give up part of their service to lagging MSs in later frames' scheduling [49]. In fact, this is a channel-aware scheduling scheme, to be further discussed in Section II-D.

Traditionally GPS assigns a fixed weight to each session. However, in applications where the absolute minimum rate and delay bound, or absolute isolation from ill-behaved sessions is not required, time-varying weight assignments can be used in order to achieve performance improvement in other criteria. When the average packet delay is concerned, the weights of the sessions can be dynamically adjusted in a way such that the average delay per packet is minimized [65]. A minimum throughput for each session is still guaranteed, while only some degree of isolation from malicious sessions is provided.

D. Channel-Aware Scheduling

In a multiple-access wireless network, the radio channel is normally characterized by time-varying fading. To exploit the characteristic, a kind of diversity (named *multiuser diversity*) can be explored to improve the system performance. The principle of multiuser diversity is that: for a cellular system with multiple MSs having independent time-varying fading channels, it is very likely that there exists an MS with the instantaneous received signal power close to its peak value. The overall resource utilization can be maximized by providing service at any time only to the MS with the best instantaneous channel quality [67].

Multiuser diversity can be applied to CDMA networks successfully. For both uplink and downlink of a CDMA system without channel fading, improved performance in system throughput can be achieved by the one-at-a-time scheduling, as indicated in Section II-A. If the time-varying channel fading is taken into account, an additional gain can be expected from a multiuser diversity scheduler, as discussed in the following. For the uplink or downlink, MS i in a cell is transmitting/receiving at instant t with E_b/I_0 value Γ_i (at the receiver side), while other MSs are idle. From (12), the achieved data rate of MS i is given by $R_i(t) = (W/\Gamma_i) \cdot \mathrm{SINR}_i(t)$, where

$$\mathrm{SINR}_i(t) = \frac{P^{\max} \cdot h_i(t)}{I_i^o(t) + \nu_i(t)} \tag{13}$$

with $P^{\max}$ being the maximum power limit of the transmitter. In the downlink with the same E_b/I_0 values, the maximum system throughput can be achieved if at any time the BS only transmits to the MS with the best instantaneous channel quality (i.e., with the highest $\mathrm{SINR}_i(t)$). For the uplink, if there is sufficient power at MS transmitters and there is also a limit on received power at the BS, when only the MS with the best channel quality transmits, the minimum transmit power is needed. This leads to minimum interference to neighbor cells, thus increasing the overall system capacity in a multicell environment.

However, the ideal multiuser diversity principle may lead to poor fairness. When the location-dependent path loss and shadowing effect are taken into account, it is very likely that only MSs near the BS and with a line-of-sight link are scheduled to transmit, while other MSs almost starve. Hence, in a practical wireless system, efforts are needed to make a good compromise between the multiuser diversity gain and fairness. Nonideal but fair multiuser diversity solutions are required, where some level of multiuser diversity gain is achieved, while at the same time each MS obtains a certain asymptotic channel access. Proportional fair scheduling algorithms are good candidates for this objective, where the transmission priority value is defined as the ratio of $\mathrm{SINR}_i(t)$ (or equivalently, the maximum allowed rate which can be supported by $\mathrm{SINR}_i(t)$) to the average throughput of MS i in a certain time window [30], [36]. If an MS has a low average throughput, its chance to be selected to transmit is relatively large, so as to alleviate the unfairness of the ideal multiuser diversity. The transmission rank can also be defined as $(\mathrm{SINR}_i(t) - \overline{\mathrm{SINR}_i})/c_i$, where $\overline{\mathrm{SINR}_i}$ is the average value of $\mathrm{SINR}_i(t)$, and c_i is a control parameter [8]. The MS with a relatively good channel has a high chance to be selected to transmit. A certain asymptotic channel access for each MS can be achieved, thus leading to fairness.

For some situations when it is better to schedule MSs simultaneously [45], [63], or when simultaneous transmissions are required by the system (e.g., wireless fair scheduling may require simultaneous transmissions, as in Section II-C), multiuser diversity benefit can still be obtained to a certain degree. In a CDMA system, the multiuser diversity can be employed more effectively and flexibly than numerous traditional channel-aware scheduling schemes for a TDMA system. For a CDMA network to implement multiuser diversity, all the MSs are divided into two sets: good channel state MSs and bad channel state MSs. The scheduler keeps track of the obtained services and channel states of all the MSs. MSs in a bad channel state postpone their transmissions until a good channel state [47], keep a relatively low power index [49], use a relatively small weight in resource allocation [65], [74], or use fewer code channels in multicode CDMA. In a good channel state, a previously sacrificed MS will be compensated, i.e., get a higher power index, a larger weight, or more code channels.

To implement multiuser diversity in a practical CDMA system, the scheduler needs to predict each MS' channel state at a future moment. The duration between the scheduling decision point and the future moment is referred to as the *prediction interval*. The multiuser diversity gain relies, to a large extent, on the accuracy of the channel prediction method. Prediction based on current and/or previous measurements of the channel can be used [47], [52]. The performance of these prediction schemes is dependent on the prediction interval and the channel coherence time. The prediction is accurate if the prediction interval is much less than the channel coherence time, and has a large prediction failure probability otherwise. In addition, if the MSs handle the channel measurement for the downlink transmission, a feedback channel is needed to report to the BS, which increases the overhead burden of the system. The induced delay in the feedback channel also degrades the prediction accuracy.

For real-time traffic with a required delay bound, if an MS is in a bad channel state for a relatively long period, its real-time traffic packets will be discarded when multiuser diversity is employed, as it has to wait until a good channel state. An effective way is to incorporate the packet delay in the scheduling decision so that the larger an MS's current packet delay, the greater its chance to be scheduled [2], [40].

In summary, in CDMA packet scheduling with QoS provisioning, a number of conflicting objectives should be taken into account, such as BER, delay, packet loss, throughput, and fairness. These factors along with the implementation complexity need to be weighted and balanced for QoS satisfaction and maximal utilization of the radio resources. The traffic flow characteristics, QoS demands, and the resource availability are key elements in determining a packet scheduling algorithm to achieve desirable system performance.

III. Network Coordination

Because of the universal frequency reuse in CDMA cellular networks, the resource allocation in each cell affects and is affected by other cells. In this context, intercell coordination can better control the resource allocation in the cells and improve overall system performance, as discussed in Section III-A and III-B, for frequency-division duplexing (FDD) and time-division duplexing (TDD) modes, respectively. In addition, to fully utilize the soft handoff, which is an inherent property of CDMA, the involved cells in a handoff should coordinate in the resource allocation to the served MSs, which is discussed in Section III-C.

A. Scheduling With Intercell Coordination in FDD

In Section II, we discuss the scheduling techniques to manage the intracell interference originating from the MS's home cell in order to achieve optimal performance in the cell. Each cell performs its own scheduling policy independent of other cells. However, due to the universal frequency reuse in CDMA systems, the transmission in one cell generates interference to the neighboring cells, which calls for the control of intercell interference. When the interference level from other cells is high, the transmission power in the target cell has to be increased to meet the E_b/I_0 requirements, thus leading to more interference to other cells. Hence, it is required to control the intercell interference in a systematic manner. One way is to resort to intercell coordination. A simple distributed intercell coordination for the uplink is to limit the total received signal power (the desired signal powers and the intercell interference) in each cell by a threshold [33]. When the intercell interference level gets higher, the total transmission power in the cell is decreased, which in turn reduces the intercell interference level. Due to the lack of possible information exchanges among the cells, this method is not able to effectively prevent possible hostile intercell interference originating from a neighboring cell.

The centralized intercell coordination can take advantage of the available traffic information in each cell and perform better scheduling to achieve better overall performance. For nonreal-time data traffic, the intercell coordination is flexible due to the delay-tolerant characteristic of data. As mentioned in Section II-A, the one-at-a-time intracell scheduling can improve system capacity for data traffic. Consider a multicell system with K cells. The BS in cell k $(1 \leq k \leq K)$ transmits only to an MS with power P_k. It is shown that each BS should either transmit at its maximum power (`on` period), or not transmit at all (`off` period) [6], in order to maximize the average per-MS throughput. So the intercell coordination is to determine when each BS is in `on` state or `off` state in order to maximize a weighted sum of all the MS's throughput values. For the optimization problem in a linear array of cells (i.e., each cell only has two neighboring cells), it is suggested that, when an MS is close to its home BS, it is served when both neighboring cells are on; and when the MS moves to one side of the cell, it is served when only the neighboring cell on the opposite side is on [6]. The general principle behind the suggestion is that an MS should not be served when it is close to an on neighboring BS. For best-effort data services with no prespecified target E_b/I_0 requirement, transmission reliability can be guaranteed by automatic repeat request (ARQ). Without the E_b/I_0 constraints, the intercell coordination is more flexible in terms of spreading gain and power allocation [56].

In a multicell environment, the effect of intercell interference can be represented by a relative intercell interference factor [68]. However, originally designed for circuit-switched and uniformly distributed voice users, this factor is not accurate enough for the future IP-based packet voice/video/data transmissions. For example, consider the following two scenarios: 1) ten data MSs randomly distributed in a cell are transmitting, each at a rate R and 2) one data MS located at the edge of a cell is transmitting at a rate $10R$. Obviously, the latter case causes more interference to the nearby neighboring cells. In general, high-rate transmission in a cellular system should be carefully controlled. This cannot be fulfilled by independent scheduling in each cell. To control the intercell interference generated from other cells, system-wise resource (i.e., power, spreading gain, and rate) allocation [41] can be employed, where an intercell coordinator takes advantage of the collected (from the BSs) traffic information and MS link status, thus achieving improved resource utilization efficiency.

Traditionally, an MS is assigned to the BS from which it receives the maximum signal strength, i.e., by the conventional least signal attenuation assignment. This BS assignment works well with a uniform distribution of the MS locations and smooth traffic. For a nonuniform MS location distribution, it is likely that one cell is overloaded, while a neighboring cell still has unused resources. Therefore, it is desired if some MSs in the overloaded cell are handed over to the neighboring cell. That is, the BS assignment should be combined with power and rate allocation, taking into account of traffic load levels in the cells [64]. When highly bursty traffic such as high-rate data is considered, the possibility of the load asymmetry among cells increases. Under this context, the system capacity can be better utilized if dynamic BS assignment is incorporated into the intercell coordination [19], [53], [64] at the cost of more computation complexity and signaling overhead.

The disadvantages of the intercell coordination mechanisms for FDD are obvious. The computation complexity is high, and some optimization problems are even NP-complete [41]. Also, high bandwidth connections between the BSs and the coordinator are required. This burden is worse with increased user mobility. As a result, suboptimal solutions may be helpful. For instance, 1) the coordinator only needs long-term channel states of the MSs, while the BSs take care of the short-term channel fluctuations, i.e., a two-tier scheduler is used [19]; 2) a small-scale optimization is performed where the optimization is only for a cluster of sectors and the dynamic BS assignment is mainly for MSs in the outer region of each cell [53]; and 3) the coordination is confined for data traffic, and the coordinator is simplified to not allowing two neighboring BSs to transmit high-rate data bursts simultaneously to their facing sectors [18].

B. Intercell Coordination in TDD-CDMA

In wireless systems, FDD works well with symmetric downlink/uplink traffic (e.g., voice). With the emerging all-IP wireless architecture and the emerging multimedia services, the downlink and uplink experience traffic load asymmetry, i.e., the downlink is characterized by a larger traffic volume due to many IP-based multimedia services such as web browsing, data downloading, and video streaming. The asymmetry level is likely to change with time as new services appear. Obviously, the FDD systems are not very flexible to handle such asymmetry, and the system capacity is limited by the up, or downlink with a relatively larger load, resulting in a waste of the bandwidth. On the other hand, TDD has the ability to effectively handle asymmetry in the downlink and uplink, and is adopted in current third–generation standards, e.g., Universal Mobile Telecommunications System (UMTS) terrestrial radio access TDD (UTRA-TDD) [26], and time-division synchronous code-division multiple access (TD-SCDMA) [16].

In a TDD system, the downlink and uplink share the same frequency band, with the BS and MSs alternately transmitting. Time is partitioned into frames of fixed duration, and each frame is divided into a fixed number of time slots. Each time slot contains a guard time (to accommodate the round trip transmission delay and the transmitting/receiving mode switching time) and can be allocated to either the downlink or the uplink. The traffic asymmetry is handled by assigning different time slot numbers to the uplink and downlink. In the same frequency band, the downlink and uplink experience a highly correlated channel fading, thus leading to channel reciprocity between them. TDD has many other merits including fast and effective open- and closed-loop power control, pre-RAKE diversity, a simple (relative to FDD) implementation of adaptive antennas, and joint detection with a reasonable complexity.

However, as the downlink and uplink of TDD share the same frequency band, each link may experience more intercell interference scenarios. Consider a specific time slot when two neighboring cells are both in downlink (or uplink) transmission. For an MS (BS) receiver in one cell, intercell interference comes from the BS (MSs) in the other cell, and is called BS-MS (MS-BS) interference. This scenario is the same as that in FDD. However, consider a specific time slot when the two neighboring cells are direction-asynchronous, e.g., cell A is in uplink and cell B is in downlink, as shown in Fig. 3. Such a time slot is called a *cross slot*. For presentation simplicity, the intracell interference is not shown in Fig. 3. In the cross slot, the BS (MS) in cell A (cell B) suffers from intercell interference from the BS (MS) in cell B (cell A), which is termed *BS-BS (MS-MS) interference*. Both BS-BS and MS-MS interference can be called *interlink interference* [76] or *same-entity interference* [27]. If the two MSs are close enough to each other (e.g., both are near the common border of the two cells), the BS-BS (MS-MS) interference can be large enough to disrupt the desired transmission. In this context, all the cells should coordinate to improve system performance.

One simple way to achieve intercell coordination and to avoid the possible service degradation due to the BS-BS or MS-MS interference is to use the same time slot configuration in all the

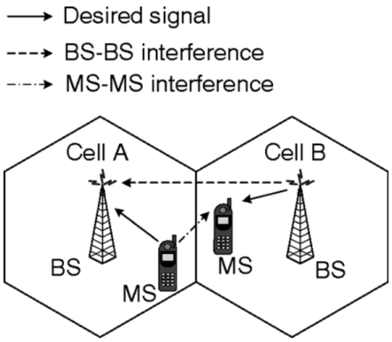


Fig. 3. Intercell interference scenario in a cross slot when cells A and B are in uplink and downlink transmission, respectively.

cells. Apparently, this may waste some portion of the capacity and lose flexibility when the traffic asymmetry level varies from cell to cell and from time to time. In order to fully utilize the capacity, dynamic allocation should be applied, where the uplink and downlink slot allocation in each cell matches closely to its traffic asymmetry level at any time. As a result, a design objective is to keep minimal BS-BS or MS-MS interference in each cross slot.

When two MSs from two neighboring cells are near the common border, the BS-BS and MS-MS interference can be significant in a cross slot. To reduce the interference, an intercell coordination mechanism called *region-based slot allocation* can be employed. If the coverage of each cell is partitioned into an inner region which is near the BS and an outer region which is far way from the BS,[2] the cross slots should be use only for uplink or downlink transmissions of the MSs in the inner region [15], [37]. For example, in Fig. 3, if the two active MSs in a cross slot are in the inner regions of cells A and B, respectively, the transmission power from/to them is not high. The two MSs are also separated by a distance. These result in insignificant BS-BS or MS-MS interference. This region-based slot allocation algorithm works well even in the worst case, where all the neighboring cells are direction-asynchronous with the cell of interest (COI). However, it is not efficient when in a cross slot some of the neighboring cells are in the same transmission direction as that of the COI, while others are not. Obviously, the slot can be assigned to MSs in the area near the common borders of the COI and the same-direction neighboring cells. Therefore, the constraint on cross slot resource allocation in the region-based algorithm can be relieved to some degree [55], which leads to higher system performance. A disjoint base station sets (DBS) method [51] can also be applied to control the possibility of large MS-MS interference. Each MS is associated with a set of the strongest BSs. If the strongest BS sets of two MSs from two cells are not disjoint, it is estimated that they may be close enough and should not share a time slot if one is transmitting and the other is receiving. The performance of this approach largely depends on the accuracy of the estimation.

The resources in TDD-CDMA can be allocated in both time domain with two transmission directions and code domain, resulting in high flexibility. However, capacity analysis and optimal resource allocation are complex. In a common uplink slot or downlink slot (among the cells of interest), the capacity of TDD remains the same as that of FDD. In a cross slot, the capacity analysis is difficult, taking into account the dynamic slot configuration among all the cells. Preliminary results have been reported for single-cell [38], two-cell [37], and multicell [15] systems. The system performance can be improved by appropriate allocation of uplink slots, downlink slots, and cross slots.

More intercell interference scenarios of TDD are traditionally viewed as a source of instability. However, these can be exploited constructively in some conditions by dynamic slot allocation [27], [28]. Consider the uplink transmission of the COI as an example. At any time slot, the intercell interference from a specific neighboring cell to the COI is either BS-BS interference (denoted I_{BB}) if the two cells are direction-asynchronous, or MS-BS interference (denoted I_{MB}) otherwise. The transmission direction of the neighboring cell at the slot can be set to downlink if $I_{\mathrm{MB}} > I_{\mathrm{BB}}$, or uplink otherwise. This can minimize the interference from the neighboring cell to the COI [28]. However, the approach opens the possibility of introducing greater interference to other neighboring cells. If the approach is applied to all the cells, the interference control solutions in different cells may conflict with each other. Therefore, it is suggested that: 1) the approach be used only in a system with a limited number of cells, e.g., when the TDD system is used only in “hot spot” traffic areas [28] and 2) the approach be carried out only under certain condition, e.g., when the transmission power of an MS exceeds a threshold [27].

In summary, the time-slot configuration in TDD-CDMA increases flexibility in resource allocation, but the complexity also increases. With the existence of cross slots, resource allocation in TDD-CDMA should be a system-wise task and should resort to intercell coordination. The capacity analysis is difficult, when the time slots are configured dynamically. In the literature, very limited results have been reported for capacity of a cross slot. More in-depth investigation is needed. Furthermore, packet scheduling and power allocation for multimedia traffic with heterogeneous QoS requirements are challenging with the existence of cross slots, therefore, requiring more sophisticated intercell coordination and deserving further research.

C. Coordination in Soft Handoff

In a wireless cellular network, an MS with an ongoing call may move away from its current cell to a neighboring cell. Occasionally, it also needs to be switched from a heavily loaded cell to a lightly loaded one for traffic load balancing. In such cases, a handoff process takes place to maintain service continuity. The future all-IP wireless networks are expected to adopt microcellular/picocellular architectures with advantages of high data rate, low-powered mobile devices, and accurate location information. In these architectures, the handoff rate will increase rapidly.

In wireless networks, an MS is in *hard handoff* if it only communicates with one BS at a time, or in *soft handoff* if it communicates with two or more BSs simultaneously. Soft handoff can eliminate “ping-pong” effect of hard handoff, leading to smooth MS communications and less signaling overhead. Soft handoff can increase the uplink capacity, thanks to macrodiversity [42]. On the other hand, its procedure is complex, and the downlink

[2] This region partitioning can be achieved by measuring the downlink pilot signal at the MSs and using a feedback channel.

macrodiversity gain is offset by the fact that transmitting from several BSs to a single MS increases interference. Although there is a small net capacity loss in the downlink soft handoff in comparison with hard handoff [46], soft handoff can offer better opportunity to transfer over the wireless link successfully. By careful planning for the degree of the overlay among cells, soft handoff can outperform hard handoff [50].

Resource allocation in soft handoff is quite different from that in hard handoff because of more coordination among neighboring cells, and is much more complicated, especially, for the downlink where the effect of soft handoff is twofold. In addition, the number of involved BSs, the traffic density, the link quality, the power distribution and power control, and the BS update rate, etc., should be considered.

For soft handoff, the set of the BSs with which an MS communicates at a time is referred to as the *active set*. Normally, the active set size is limited by a maximum number. For neighboring BSs to join/leave the active set, the second-generation IS-95A uses a static threshold algorithm. Each MS monitors its received pilot signals (from BSs). If the received strength of a pilot from a BS exceeds a threshold T_ADD and the active set size is less than the maximum value, the BS is included into the active set of the MS. When the strength of the pilot from the BS drops below a threshold T_DROP (T_DROP $<$ D_ADD), the BS is removed from the MS's active set. This algorithm is simple and easy to implement. Each BS in the active set is expected to provide a transmission to the MS with reasonable quality. However, the algorithm does not consider the variation (among neighboring cells) in traffic density, path quality, and interference level. It may lead to inefficiency in resource allocation. Hence, it is desired to dynamically adjust the thresholds to control the resource usage and transmission quality [31], [60]. A heuristic algorithm can allow the two thresholds to vary based on traffic density of the cells. Specifically, the T_ADD (T_DROP) can take a different value in a heavy traffic load case from that in a light traffic load case [34]. A dynamic threshold algorithm is also adopted in the IS-95B/CDMA2000. Dynamical thresholds can lead to a certain gain in system capacity, at the cost of the possibility of frequent BS update in the active set.

In downlink soft handoff, several transmissions are kept for each MS. A conventional implementation is to let multiple BSs in the active set transmit with the same power level to the MS and apply the same power correction (according to the power control command sent by the MS). Not considering the link condition variation in different BSs, the above *equal power distribution* may lead to a significant capacity loss [43], [46], although it is simple to implement. Therefore, the transmission power control and distribution need to be redesigned in order to improve system capacity [81]. The power control scheme should be differentiated among the BSs by adapting to the radio link gain, interference level, and total transmitted power in each cell [12]. Another implementation of power control differentiation is site selection diversity transmission power control (SSDT) [23]. In SSDT, a BS in the active set is selected dynamically at a time as the primary BS, and transmits to the MS with adequate power, while the output power of other BSs in the active set is held to a minimum level. This can mitigate the interference caused by multiple transmissions from several BSs with conventional transmission power control. In addition, an appropriate power distribution among the BSs in the active set is necessary. For an MS i in soft handoff, let $\mathcal{S}_i$ denote the active set. With maximum ratio combining, the actual SINR at the MS receiver output with coherent detection is given by $\gamma_i = \sum_{j \in \mathcal{S}_i} \gamma_{ij}$, where γ_{ij} is the SINR for the received signal from BS j in the active set. Let γ_i^* be the desired SINR value for MS i, and $\xi_{ij} = \gamma_{ij}/\gamma_i^*$ a parameter to adjust the power distribution for the transmission from BS j. Then, the QoS requirement $\gamma_i = \gamma_i^*$ is equivalent to $\sum_{j \in \mathcal{S}_i} \xi_{ij} = 1$. For B BSs and N MSs, the matrix $\boldsymbol{\xi} = \{\xi_{ij}, i = 1, \ldots, N, j = 1, \ldots, B\}$ is referred to as the *power distribution association matrix*, where $\xi_{ij} = 0$ if BS j is not in the active set of MS i. To achieve the best system performance, it is desired to find the optimal $\boldsymbol{\xi}$ which minimizes the transmission outage probability (defined as the probability that an MS cannot get a guaranteed service in a transmission slot) [80].

The performance of a soft handoff algorithm can be measured by the metrics of active set size, active set update rate, and signal quality. Intuitively, a large active set can achieve more macrodiversity gain and improve the signal quality, at the cost of more system resource consumption. If a low active set size is maintained, the active set update (or handoff) rate should be increased in order to keep a certain level of signal quality, at the expense of update overhead and switching cost. Therefore, a design tradeoff should be balanced carefully among the three metrics [5]. Normally, optimal soft handoff algorithms need a complicated trajectory model of the MS which may not be available and/or accurate in practice. In this context, a locally optimal handoff algorithm may be more practical, where the motion of the MS can be approximated by a straight line as a small time-scale is adopted [60].

In a packet-switched mode or in a TDD mode, it is much more complex to achieve soft handoff as additional transmission time-slot synchronization is required among the transmissions between the BSs and the MS. Therefore, handoff in current UTRA-TDD standard usually means hard handoff [26]. To implement soft handoff in these networks, further research is necessary.

IV. Further Discussion and Conclusion

Future wireless networks are evolving toward an all-IP heterogeneous architecture, which includes different wireless networks to provide seamless Internet access to MSs [54]. Recently, the differentiated services (DiffServ) approach has emerged as an scalable solution to ensure QoS in IP networks. For the integration of heterogeneous all-IP wireless networks with the Internet backbone, a domain-based DiffServ architecture provides a flexible and efficient solution, because each DiffServ domain can independently select, modify, or exchange its own internal resource management mechanism to implement its service level agreements (SLAs) with neighboring domains [17]. Many challenges are posed when applying DiffServ over CDMA wireless domains.

- An effective resource management is necessary to simultaneously provide DiffServ QoS and achieve efficient resource utilization in the low layers, taking into account the hostile propagation environment, user mobility, and interference-limited capacity. The achieved DiffServ QoS over CDMA networks should be determined by the mechanisms in both the IP layer and the link/physical layers.

- When a data application transported by TCP is provided with DiffServ QoS, there exists an interaction between the TCP congestion management and the IP-layer traffic conditioning/forwarding mechanism [25]. When DiffServ is applied to CDMA wireless networks, this interaction should be extended to the link/physical layers.
- Effective packet-switching based CAC is required. In IP-based CDMA networks, the traffic arrival is bursty. Due to user mobility, bursty traffic patterns and the various transmission accuracy requirements (over the wireless channel) of multimedia traffic, the CDMA system capacity varies with time even in a nonfading scenario. It becomes more dynamic for the TDD mode due to the more complex intercell interference scenarios.

On the other hand, cross-layer design has emerged as a promising solution to meet the various QoS requirements of multimedia traffic over wireless links. Better performance can be obtained from the information exchanges across the protocol layers which may not be available in the traditional layering architecture. The channel-aware scheduling discussed in Section II-D is actually a cross-layer design approach (cross-layer between the link layer and physical layer). In addition, two other cross-layer design approaches exist for IP-based CDMA cellular networks.

- TCP over CDMA links: When a TCP connection is served by CDMA wireless channels, TCP interacts with the CDMA link-layer resource allocation. Specifically, TCP dynamically adjusts the sending rate of TCP segments (which will be fed into the link-layer transmission queue) according to network congestion status (e.g., packet loss and round-trip delay); on the other hand, the link-layer resource allocation (e.g., power, time slot and rate) ultimately determines the packet loss rate and transmission delay over the wireless link, and therefore, affects the TCP performance. Taking into account the interaction, a cross-layer design approach is expected to improve the overall system performance. Although some preliminary results have been reported [39], more research efforts need to be made in this area.
- Joint source/channel coding—power/rate allocation for video services: An effective way for video transmission over wireless links is to use joint source-channel coding [20]. For CDMA networks with interference-limited capacity, it is important to take into account the power/rate management in CDMA systems when designing the source and channel coding. More flexibility can be obtained when power/rate allocation is considered jointly with the source and/or channel coding. However, the performance optimization is also more complicated to achieve [13], [21], [78]. The case is worse when time scheduling for multiplexed video traffic is implemented. Further investigation is necessary.

To summarize, in this paper, the fundamentals of recent research efforts on CDMA-related QoS provisioning with efficient resource utilization in future cellular networks have been presented. Packet scheduling and power allocation have the ability to guarantee bit-level and packet-level QoS, under the interference-limited system capacity. Network coordination is effective to achieve performance gains, at the cost of additional complexity and signaling overhead. Other research issues include the suboptimal intercell coordination with acceptable computation complexity and signaling overhead, capacity analysis for TDD-CDMA, efficient resource allocation for multimedia traffic with heterogeneous QoS requirements in TDD-CDMA with a dynamic slot configuration, effective soft handoff in packet-switched CDMA networks, DiffServ QoS over CDMA networks, and effective cross-layer design to achieve overall system performance improvement.

References

[1] I. F. Akyildiz, D. A. Levine, and I. Joe, "A slotted CDMA protocol with BER scheduling for wireless multimedia networks," *IEEE/ACM Trans. Netw.*, vol. 7, no. 2, pp. 146–158, Apr. 1999.

[2] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, Feb. 2001.

[3] M. A. Arad and A. Leon-Garcia, "Scheduled CDMA: A hybrid multiple access for wireless ATM networks," in *Proc. IEEE PIMRC*, 1996, pp. 913–917.

[4] M. A. Arad and A. Leon-Garcia, "A generalized processor sharing approach to time scheduling in hybrid CDMA/TDMA," in *Proc. IEEE INFOCOM*, 1998, pp. 1164–1171.

[5] M. Asawa and W. E. Stark, "Optimal scheduling of handoffs in cellular networks," *IEEE/ACM Trans. Netw.*, vol. 4, no. 3, pp. 428–441, Jun. 1996.

[6] A. Bedekar, S. C. Borst, K. Ramanan, P. A. Whiting, and E. M. Yeh, "Downlink scheduling in CDMA data networks," Lucent Technologies, Tech. Memo., 1999.

[7] J. C. R. Bennett and H. Zhang, "$\mathrm{WF}^2\mathrm{Q}$: Worst-case fair weighted fair queueing," in *Proc. IEEE INFOCOM*, 1996, pp. 120–128.

[8] F. Berggren and R. Jantti, "Asymptotically fair transmission scheduling over fading channels," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 326–336, Jan. 2004.

[9] F. Berggren, S.-L. Kim, R. Jantti, and J. Zander, "Joint power control and intracell scheduling of DS-CDMA nonreal time data," *IEEE J. Select. Areas Commun.*, vol. 19, no. 10, pp. 1860–1870, Oct. 2001.

[10] F. Berggren and S.-L. Kim, "Energy-efficient control of rate and power in DS-CDMA systems," *IEEE Trans. Wireless Commun.*, vol. 3, no. 3, pp. 725–733, May 2004.

[11] Q. Bi and J. Seymour, "The future evolution of wireless mobile communications," *Wireless Commun. Mobile Comput.*, vol. 3, no. 6, pp. 705–716, Sep. 2003.

[12] F. Blaise, L. Elicegui, F. Goeusse, and G. Vivier, "Power control algorithms for soft handoff users in UMTS," in *Proc. IEEE Veh. Technol. Conf.*, 2002, pp. 1110–1114.

[13] Y. S. Chan and J. W. Modestino, "A joint source coding-power control approach for video transmission over CDMA networks," *IEEE J. Select. Areas Commun.*, vol. 21, no. 10, pp. 1516–1525, Dec. 2003.

[14] C.-S. Chang and K.-C. Chen, "Medium access protocol design for delay-guaranteed multicode CDMA multimedia networks," *IEEE Trans. Wireless Commun.*, vol. 2, no. 6, pp. 1159–1167, Nov. 2003.

[15] K.-N. Chang and K.-D. Lee, "Capacity of multicell CDMA/shared-TDD system," in *Proc. IEEE PIMRC*, 2003, pp. 891–895.

[16] H.-H. Chen, C.-X. Fan, and W. W. Lu, "China's perspectives on 3G mobile communications and beyond: TD-SCDMA technology," *IEEE Wireless Commun.*, vol. 9, no. 2, pp. 48–59, Apr. 2002.

[17] Y. Cheng, H. Jiang, W. Zhuang, Z. Niu, and C. Lin, "Efficient resource allocation for China's 3G/4G wireless networks," *IEEE Commun. Mag.*, vol. 43, no. 1, pp. 76–83, Jan. 2005.

[18] S. S. Choi and D. H. Cho, "Coordinated resource allocation scheme for forward link in sectorized CDMA systems," in *Proc. IEEE VTC Fall*, 2002, pp. 2356–2360.

[19] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *Proc. IEEE INFOCOM*, 2003, pp. 786–796.

[20] C. Dubuc, D. Boudreau, and F. Patenaude, "The design and simulated performance of a mobile video telephony application for satellite third-generation wireless systems," *IEEE Trans. Multimedia*, vol. 3, no. 4, pp. 424–431, Dec. 2001.

[21] Y. Eisenberg, C. E. Luna, T. N. Pappas, R. Berry, and A. K. Katsaggelos, "Joint source coding and transmission power management for energy efficient wireless video communications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 411–424, Jun. 2002.

[22] Y. Fang and Y. Zhang, "Call admission control schemes and performance analysis in wireless mobile networks," *IEEE Trans. Veh. Technol.*, vol. 51, no. 2, pp. 371–382, Mar. 2002.

[23] H. Furukawa, K. Hambe, and A. Ushirokawa, "SSDT-site selection diversity transmission power control for CDMA forward link," *IEEE J. Select. Areas Commun.*, vol. 18, no. 8, pp. 1546–1554, Aug. 2000.

[24] L. Georgiadis, R. Guerin, and A. Parekh, "Optimal multiplexing on a single link: Delay and buffer requirements," *IEEE Trans. Inf. Theory*, vol. 43, no. 5, pp. 1518–1535, Sep. 1997.

[25] P. Giacomazzi, L. Musumeci, and G. Verticale, "Transport of TCP/IP traffic over assured forwarding IP-differentiated services," *IEEE Network*, vol. 17, no. 5, pp. 18–28, Sep.-Oct. 2003.

[26] M. Haardt, A. Klein, R. Koehn, S. Oestreich, M. Purat, V. Sommer, and T. Ulrich, "The TD-CDMA based UTRA TDD mode," *IEEE J. Select. Areas Commun.*, vol. 18, no. 8, pp. 1375–1385, Aug. 2000.

[27] H. Haas and S. McLaughlin, "A dynamic channel assignment algorithm for a hybrid TDMA/CDMA-TDD interface using the novel TS-opposing technique," *IEEE J. Select. Areas Commun.*, vol. 19, no. 10, pp. 1831–1846, Oct. 2001.

[28] H. Haas and S. McLaughlin, "A novel channel assignment approach in TDMA/CDMA-TDD systems," in *Proc. IEEE PIMRC*, 2001, pp. E-142–E-146.

[29] H. Holma and A. Toskala, Eds., *WCDMA for UMTS*. New York: Wiley, 2000.

[30] J. M. Holtzman, "Asymptotic analysis of proportional fair algorithm," in *Proc. IEEE PIMRC*, 2001, pp. F-33–F-37.

[31] B. Homnan and W. Benjapolakul, "QoS-controlling soft handoff based on simple step control and a fuzzy inference system with the gradient descent method," *IEEE Trans. Veh. Technol.*, vol. 53, no. 3, pp. 820–834, May 2004.

[32] V. Huang and W. Zhuang, "QoS-oriented packet scheduling for wireless multimedia CDMA communications," *IEEE Trans. Mobile Comput.*, vol. 3, no. 1, pp. 73–85, Jan.-Mar. 2004.

[33] V. Huang and W. Zhuang, "QoS based fair resource allocation in multicell TD/CDMA communication systems," *IEEE Trans. Wireless Commun.*, to be published.

[34] S.-H. Hwang, S.-L. Kim, H.-S. Oh, C.-E. Kang, and J.-Y. Son, "Soft handoff algorithm with variable thresholds in CDMA cellular systems," *Electron. Lett.*, vol. 33, no. 19, pp. 1602–1603, Sep. 1997.

[35] S. A. Jafar and A. Goldsmith, "Optimal rate and power adaptation for multirate CDMA," in *Proc. IEEE Veh. Technol. Conf., Fall*, 2000, pp. 994–1000.

[36] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. IEEE Veh. Technol. Conf., Spring*, 2000, pp. 1854–1858.

[37] W. S. Jeon and D. G. Jeong, "Comparison of time slot allocation strategies for CDMA/TDD systems," *IEEE J. Select. Areas Commun.*, vol. 18, no. 7, pp. 1271–1278, Jul. 2000.

[38] D. G. Jeong and W. S. Jeon, "CDMA/TDD system for wireless multimedia services with traffic unbalance between uplink and downlink," *IEEE J. Select. Areas Commun.*, vol. 17, no. 5, pp. 939–946, May 1999.

[39] H. Jiang and W. Zhuang, "Cross-layer resource allocation for integrated voice/data traffic in wireless cellular networks," *IEEE Trans. Wireless Commun.*, to be published.

[40] H. Jiang and W. Zhuang, "Resource allocation with service differentiation for wireless video transmission," *IEEE Trans. Wireless Commun.*, to be published.

[41] D. I. Kim, E. Hossain, and V. K. Bhargava, "Downlink joint rate and power allocation in cellular multirate WCDMA systems," *IEEE Trans. Wireless Commun.*, vol. 2, no. 1, pp. 69–80, Jan. 2003.

[42] J. Y. Kim, G. L. Stüber, and I. F. Akyildiz, "A simple performance/capacity analysis of multiclass macrodiversity CDMA cellular systems," *IEEE Trans. Commun.*, vol. 50, no. 2, pp. 304–308, Feb. 2002.

[43] J. Y. Kim and G. L. Stüber, "CDMA soft handoff analysis in the presence of power control error and shadowing correlation," *IEEE Trans. Wireless Commun.*, vol. 1, no. 2, pp. 245–255, Apr. 2002.

[44] K. Kumaran and L. Qian, "Scheduling on uplink of CDMA packet data network with successive interference cancellation," in *Proc. IEEE WCNC*, 2003, pp. 1645–1650.

[45] K. Kumaran and L. Qian, "Uplink scheduling in CDMA packet-data systems," in *Proc. IEEE INFOCOM*, 2003, pp. 292–300.

[46] C.-C. Lee and R. Steele, "Effect of soft and softer handoffs on CDMA system capacity," *IEEE Trans. Veh. Technol.*, vol. 47, no. 3, pp. 830–841, Aug. 1998.

[47] D.-J. Lee and D.-H. Cho, "Data transmission scheduling considering short-term fading for transmit power reduction in CDMA systems," *IEEE Trans. Veh. Technol.*, vol. 51, no. 6, pp. 1621–1627, Nov. 2002.

[48] C. W. Leong, W. Zhuang, Y. Cheng, and L. Wang, "Call admission control for wireless systems supporting integrated on/off voice and best effort data services," *IEEE Trans. Commun.*, vol. 52, no. 5, pp. 778–790, May 2004.

[49] C. Li and S. Papavassiliou, "Fair channel-adaptive rate scheduling in wireless networks with multirate multimedia services," *IEEE J. Select. Areas Commun.*, vol. 21, no. 10, pp. 1604–1614, Dec. 2003.

[50] Y.-B. Lin and A.-C. Pang, "Comparing soft and hard handoffs," *IEEE Trans. Veh. Technol.*, vol. 49, no. 3, pp. 792–798, May 2000.

[51] M. Lindstrom, "Improved TDD resource allocation through inter-mobile interference avoidance," in *Proc. IEEE Veh. Technol. Conf., Spring*, 2001, pp. 1027–1031.

[52] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Trans. Netw.*, vol. 7, no. 4, pp. 473–489, Aug. 1999.

[53] C. Makaya and S. Aissa, "Joint scheduling and base station assignment for CDMA packet data networks," in *Proc. IEEE Veh. Technol. Conf., Fall*, 2003, pp. 1693–1697.

[54] J. W. Mark and W. Zhuang, *Wireless Communications and Networking*. Upper Saddle River, NJ: Prentice-Hall, 2003.

[55] J. Nasreddine and X. Lagrange, "Time slot allocation based on a path gain division scheme for TD-CDMA TDD systems," in *Proc. IEEE Veh. Technol. Conf., Spring*, 2003, pp. 1410–1414.

[56] S.-J. Oh, T. L. Olsen, and K. M. Wasserman, "Distributed power control and spreading gain allocation in CDMA data networks," in *Proc. IEEE INFOCOM*, 2000, pp. 379–385.

[57] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single-node case," *IEEE/ACM Trans. Netw.*, vol. 1, no. 3, pp. 344–357, Jun. 1993.

[58] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The multiple node case," *IEEE/ACM Trans. Netw.*, vol. 2, no. 2, pp. 137–150, Apr. 1994.

[59] J. Perez-Romero, O. Sallent, and R. Agusti, "Traffic and physical layer effects on packet scheduling design in W-CDMA systems," *Electron. Lett.*, vol. 38, no. 16, pp. 917–918, Aug. 2002.

[60] R. Prakash and V. V. Veeravalli, "Locally optimal soft handoff algorithms," *IEEE Trans. Veh. Technol.*, vol. 52, no. 2, pp. 347–356, Mar. 2003.

[61] S. Ramakrishna and J. M. Holtzman, "A scheme for throughput maximization in a dual-class CDMA system," *IEEE J. Select. Areas Commun.*, vol. 16, no. 6, pp. 830–844, Aug. 1998.

[62] A. Sampath, P. Sarath Kumar, and J. M. Holtzman, "Power control and resource management for a multimedia CDMA wireless system," in *Proc. IEEE PIMRC*, 1995, pp. 21–25.

[63] T. Shu and Z. Niu, "A channel-adaptive and throughput-efficient scheduling scheme in voice/data DS-CDMA networks with constrained transmission power," in *Proc. IEEE ICC*, 2003, pp. 2229–2233.

[64] M. Soleimanipour, W. Zhuang, and G. H. Freeman, "Optimal resource management in wireless multimedia wideband CDMA systems," *IEEE Trans. Mobile Comput.*, vol. 1, no. 2, pp. 143–160, Apr.–Jun. 2002.

[65] A. Stamoulis, N. D. Sidiropoulos, and G. B. Giannakis, "Time-varying fair queueing scheduling for multicode CDMA based on dynamic programming," *IEEE Trans. Wireless Commun.*, vol. 3, no. 2, pp. 512–523, Mar. 2004.

[66] A. C. Varsou, H. C. Huang, and L. Mailaender, "Rate scheduling for the downlink of CDMA mixed traffic networks," in *Proc. IEEE WCNC*, 2000, pp. 370–375.

[67] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1277–1294, Jun. 2002.

[68] A. J. Viterbi, A. M. Viterbi, and E. Zehavi, "Other-cell interference in cellular power-controlled CDMA," *IEEE Trans. Commun.*, vol. 42, no. 2,3,4, pp. 1501–1504, Feb./Mar./Apr. 1994.

[69] L. Wang and W. Zhuang, "Call admission control for packet data in CDMA cellular communications," *IEEE Trans. Wireless Commun.*, to be published.

[70] X. Wang, "An FDD wideband CDMA MAC protocol for wireless multimedia networks," in *Proc. IEEE INFOCOM*, 2003, pp. 734–744.

[71] ——, "Wide-band TD-CDMA MAC with minimum-power allocation and rate- and BER-scheduling for wireless multimedia networks," *IEEE/ACM Trans. Netw.*, vol. 12, no. 1, pp. 103–116, Feb. 2004.

[72] J.-H. Wen, J.-K. Lain, and Y.-W. Lai, "Performance evaluation of a joint CDMA/NC-PRMA protocol for wireless multimedia communications," *IEEE J. Select. Areas Commun.*, vol. 19, no. 1, pp. 95–106, Jan. 2001.

[73] L. Xu, X. Shen, and J. W. Mark, "Dynamic fair scheduling with QoS constraints in multimedia wideband CDMA cellular networks," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 60–73, Jan. 2004.

[74] L. Xu, X. Shen, and J. W. Mark, "Fair resource allocation with guaranteed statistical QoS for multimedia traffic in wideband CDMA cellular network," *IEEE Trans. Mobile Comput.*, vol. 4, no. 2, pp. 166–177, Mar.–Apr. 2005.

[75] J. Ye, X. Shen, and J. W. Mark, "Call admission control in wideband CDMA cellular networks by using fuzzy logic," *IEEE Trans. Mobile Comput.*, vol. 4, no. 2, pp. 129–141, Mar.-Apr. 2005.

[76] H. Yomo and S. Hara, "An uplink/downlink asymmetric slot allocation algorithm in CDMA/TDD-based wireless multimedia communications systems," in *Proc. IEEE Veh. Technol. Conf., Fall*, 2001, pp. 797–801.

[77] J.-H. Yoon, M.-J. Sheen, and S.-C. Park, "Scheduling methods with transmit power constraint for CDMA packet services," in *Proc. IEEE Veh. Technol. Conf., Spring*, 2003, pp. 1450–1453.

[78] Q. Zhang, Z. Ji, W. Zhu, and Y.-Q. Zhang, "Power-minimized bit allocation for video communication over wireless channels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 398–410, Jun. 2002.

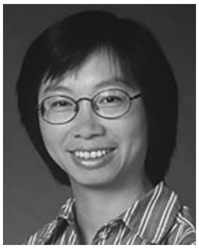
[79] D. Zhao, X. Shen, and J. W. Mark, "Radio resource management for cellular CDMA systems supporting heterogeneous services," *IEEE Trans. Mobile Comput.*, vol. 2, no. 2, pp. 147–160, Apr.–Jun. 2003.

[80] D. Zhao, X. Shen, and J. W. Mark, "QoS guarantee and power distribution for soft handoff connections in cellular CDMA downlinks," *IEEE Trans. Wireless Commun.*, to be published.

[81] D. Zhao, X. Shen, and J. W. Mark, "Soft handoff and connection reliability in cellular CDMA downlink," *IEEE Trans. Wireless Commun.*, to be published.

Hai Jiang (S'04) received the B.S. and M.S. degrees from Peking University, Beijing, China, both in electronics engineering, in 1995 and 1998, respectively. He is currently working towards the Ph.D. degree at the Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada.

His current research interests include quality-of-service provisioning and resource management for multimedia communications in all-IP wireless networks.



Weihua Zhuang (M'93–SM'01) has been a Professor in the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, since 1993. She is a coauthor of the textbook *Wireless Communications and Networking* (Prentice-Hall, 2003). Her current research interests include multimedia wireless communications, wireless networks, and radio positioning.

Dr. Zhuang received the Outstanding Performance Award in 2005 from the University of Waterloo and the Premier's Research Excellence Award (PREA) in 2001 from the Ontario Government. She is an Editor/Associate Editor of the IEEE Transactions on Wireless Communications, the IEEE Transactions on Vehicular Technology, and the *EURASIP Journal on Wireless Communications and Networking*.

Xuemin (Sherman) Shen (M'97–SM'02) has been with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, since October 1993, where he is a Professor and the Associate Chair for Graduate Studies. He is a coauthor of two books. His research focuses on mobility and resource management in interconnected wireless/wireline networks, ultra-wideband (UWB) wireless communications systems, wireless security, and ad hoc and sensor networks.

Dr. Shen received the Outstanding Performance Award in 2004 from the University of Waterloo, the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, and the Distinguished Performance Award in 2002 from the Faculty of Engineering, University of Waterloo. He serves on many Editorial Boards including the IEEE Transactions on Wireless Communications, the IEEE Transactions on Vehicular Technology, *ACM/Wireless Networks*, and *Computer Networks*.

Qi Bi (SM'92) received the B.S. and M.S. degrees from Shanghai Jiao Tong University, Shanghai, China, and the Ph.D. from Pennsylvania State University, University Park.

He is a Bell Laboratories Fellow in the Mobility Solutions Unit, System Engineering Department, Lucent Technologies, Whippany, NJ. He currently heads a team with responsibilities of analyzing and designing the third-generation wireless digital communication systems. He served as the Guest Editor of *Wireless Communications and Mobile Computing* (Wiley). He is also a recognized leader outside of Lucent Technologies and has served as technical chair in many international conferences. He holds more than 40 U.S. patents. His present focus is in the areas of high-speed wireless data network delivering VoIP, broadcast and multicast services, push to talk, and broadband wireless communications.

Dr. Bi was the recipient of numerous honors including the Advanced Technology Laboratory Award in 1995 and 1996, the Bell Laboratories President's Gold Award in 2000 and 2002, The Bell Laboratories Innovation Team Award in 2003, the Speaker of the Year Award from the IEEE New Jersey Coast Section in 2004, and the Asian American Engineer of the Year Award in 2005. He has served as the Technical Vice-Chair of the IEEE Wireless Communications and Network Conference 2003, Technical Chair for Wireless Symposium of the IEEE GLOBECOM 2000–2002, and organizer of the First and Second Lucent IS-95 and UMTS Technical Conference in 1999 and 2000. He served as Feature Editor of the *IEEE Communications Magazine* (2001), Editor of the IEEE Journal on Selected Areas in Communications and the IEEE Transaction on Wireless Communications.