Joint Optimization of Hybrid ARQ Performance in Wireless Packet Data Systems

Jeong-woo Cho, Member, IEEE, Jeonghoon Mo, Member, IEEE, and Yong Chang, Member, IEEE

Abstract

We propose an analytical framework that optimizes the performance of various Hybrid Automatic Repeat reQuest (HARQ) schemes in wireless packet data systems. The optimization framework lends itself to maximizing the user throughput while the redundancy size per retransmission, the packet error rate, and the mapping of signal to interference plus noise ratios (SINR) into modulation and coding scheme (MCS) levels are jointly decided. After deriving a succinct throughput expression covering incremental redundancy (IR) scheme as well as Chase combining (CC) scheme, we show that there exists an optimal redundancy size for IR scheme when the number of retransmissions is limited, while on the other hand as small as possible redundancy size is optimal when the number of retransmissions is unlimited. Through extensive simulations of the IEEE 802.16e system which verify our analytical results, it is shown that IR scheme significantly outperforms CC scheme even with a practical packet error rate constraint.

Index Terms

Cross-layer design and optimization, Hybrid automatic repeat request, incremental redundancy, optimization, IEEE 802.16e system

I. INTRODUCTION

Even with untiring efforts to satisfy service requirements of various applications, wireless packet data systems have not yet matched wired networks, e.g., the Internet, in throughput. Since the ultimate demand of mobile users is to enjoy high data throughput comparable to that of wired networks, wireless packet data systems have been evolving into more advanced systems such as the evolved UTRA [1] and IEEE 802.16e [2] where Hybrid Automatic Repeat reQuest (HARQ) schemes as well as broader frequency bands and orthogonal frequency division multiple access (OFDMA) have been the essentials to their performance.

A preliminary version of this paper is to be presented at the IEEE Globecom 2006, San Francisco, CA.
Jeong-woo Cho and Yong Chang are Senior Engineers at Samsung Electronics, Suwon, Korea (e-mail: ggumdol@gmail.com).
Jeonghoon Mo is with the School of Engineering, Information and Communications University, Daejeon, Korea (e-mail: jhmo@icu.ac.kr).
Unlike the simple ARQ scheme where a receiver scraps former erroneous packets and tries to decode each packet independently, an HARQ receiver buffers erroneous packets until it successfully decodes the original data by exploiting the signal information of the buffered packets. Since the soft information of symbols in the initial transmission are combined with that of the following retransmissions, each retransmission increases the probability of successful decoding at the receiver.

There are two fundamental characteristics of HARQ contributing to the performance improvement of systems adopting HARQ. That is to say, HARQ makes decoding attempts in 1) combinative and 2) iterative ways. Firstly, an HARQ receiver combines the signal information of packets received so far and that of the last packet to form an aggregate signal information. If an HARQ receiver does not exploit the combined information, HARQ might be reduced to simple ARQ. Secondly, if the HARQ receiver does not try to decode per packet reception, air resources are wasted since more than enough packets might be transmitted. In this case, HARQ is reduced to a simple forward error correction (FEC) scheme since it does not utilize the early termination probability.

To utilize these characteristics better, various HARQ schemes have been introduced. In Chase combining (CC) scheme [3], a sender retransmits the original packet for each decoding failure of its corresponding receiver until the receiver successfully decodes the combined information. Thus the same packet is used for all retransmission requests in CC scheme. In incremental redundancy (IR) scheme [4], [5], a sender encodes the original (systematic) bits to produce the highest rate code of the corresponding code family. The sender transmits only the systematic bits at first and transmits one redundancy block for each retransmission failure of its corresponding receiver. In contrast to CC, a redundancy block is not necessarily the original packet but can be one of encoded blocks and the redundancy block size can be finely adjusted so that an appropriate amount of redundancy bits is delivered to the receiver. As the receiver combines more redundancy blocks with the original packet, the effective code rate becomes lowered to achieve a higher coding gain. In general, Turbo codes [4], Rate Compatible Punctured Convolutional (RCPC) [5] codes, and Low-Density Parity-Check (LDPC) [6] codes are used to generate redundancy blocks. Note that CC is a special case of IR and is thus implicitly supported in wireless packet data systems adopting IR scheme, e.g. [1] and [2].

II. MOTIVATION AND RELATED WORKS

To improve the actual performance of user applications, the following performance indices should be improved: user throughput, packet error rate (PER), and packet delay where packet delay and packet error rate are closely related to each other. To better understand the importance of the performance indices, let us consider two representative applications one by one. A real-time application generally requires a lower packet delay which is achieved by controlling the packet error rate and the number of retransmissions. For
a real-time application, packets received after a certain number of retransmissions is useless as packets should arrive timely within a specified time bound. Moreover, it is well known that transport layer protocols such as TCP NewReno and TCP Reno, that are mainly adopted by delay-tolerant applications, are very vulnerable to even low packet error rates and large packet delays [7]. The higher the packet error rate or the larger the packet delay, the lower the TCP throughput. Therefore, achieving sufficiently low packet error rates is crucial for both applications. Furthermore, it is certain that an advanced wireless packet data system should also provision abundant user throughputs to satisfy both applications. We can conclude that all of three performance indices are indispensable for satisfying the performance demand of user applications.

Though there have been a few approaches to improve the performance of HARQ from different angles, against our expectations, we have found that there has not been a unified approach that maximizes the throughput performance of HARQ while simultaneously maintaining the packet error rate and the packet delay at sufficiently low values. In almost work including [8] and [9], either 1) only CC scheme was analyzed or 2) IR scheme was analyzed without adjusting the redundancy block size or 3) the mapping of signal to interference plus noise ratios (SINR) into modulation and coding scheme (MCS) levels was simply determined such that the packet error rate is less than or equal to a certain threshold. Unlike [8] and [9], Zheng and Viswanathan in their noticeable work [10], introduced a systematic and analytical approach that maximizes the user throughput of CC scheme in downlink transmissions of wireless packet data systems and proposed several scheduling algorithms. Though their framework is thought-provoking from diverse standpoints of HARQ optimization, they neither did consider IR schemes nor did impose a constraint upon the packet error rate.

In this paper, we provide an extended analytical framework which utilizes the inherent capability of IR scheme, i.e., adjusting the redundancy block size. That is to say, we would like to answer the following questions in particular:

*How much improvement can we make by optimally adjusting the redundancy block size of IR scheme?*

To answer this question, we formulate an optimization problem which maximizes the user throughput in wireless packet data systems by jointly determining the redundancy size per retransmission, the packet error rate, and the mapping of SINR into MCS levels. In addition, a practical constraint is imposed on the packet error rate so that the packet delay and retransmission probability are kept under control. We also derive a concise expression of the user throughput and analyze the asymptotic behavior of HARQ schemes as the number of retransmissions is limited and goes to infinity, that were not addressed in previous works, either. Finally, we show through extensive simulations that our optimization approach with variable redundancy block size significantly improves the user throughput of an existing system, i.e.,
IEEE 802.16e system, while maintaining a designated residual packet error rate.

III. SYSTEM MODEL

We consider a downlink packet transmission with an HARQ scheme from a base station (BS) to a mobile station (MS) with a lossy channel in which packets can be corrupted with some probability in an OFDMA based system. The BS transmits a block of size $B$ bits at a data rate $r_1$ which is determined by the selected modulation and coding scheme (MCS) level. Upon a successful transmission, the BS completes the transmission and prepares the next block while upon the failure, it retransmits either the original block or the redundant block, where the retransmitted block size at the $k$-th trial is $\alpha_k \cdot B$ for $k = 1, 2, \ldots, \bar{k}$ and $\alpha_k$ is a value greater than 0. The retransmission is repeated either the number of trials including the first transmission reaches $\bar{k}$ (with a possibility of $\bar{k} = \infty$), or the packet is successfully delivered. If the number of trials is more than $\bar{k}$, the BS relies on a higher layer to recover packets.

There are $L$ available MCS levels, $\{(m_l, c_l), l = 1, 2, \ldots, L\}$, to the BS where $m_l$ and $c_l$ represent modulation and coding of level $l$, respectively. We assume that the transmission rate $mc_l := m_l \times c_l$ is sorted in an ascending order, i.e., $mc_1 \leq mc_2 \leq \cdots \leq mc_L$. The BS can adaptively change its MCS and selects one of them based on the channel condition represented by signal to interference plus noise ratio (SINR) $\gamma$. We assume that the BS is aware of SINR in the model and $\gamma_k$ designates the SINR of the channel at the $k$-th trial. In the existing wireless packet data systems, the BS selects its MCS level $l_k^*$ of the $k$-th trial such that the actual data rate $r_k$ is maximized:

$$l_k^* = \arg \max_{l = 1, 2, \ldots, L} \left\{ r_k = mc_l \mid P_e(r_k, \gamma_k) \leq P_T \right\} ,$$

where $P_e(r, \gamma)$ is the loss rate for a given data rate $r$ and SINR $\gamma$. In other words, it selects the highest data rate as long as the error rate is less than the threshold $P_T$.

To return to the subject, upon the reception of the $k$-th block, the receiver decodes the transmitted blocks and combines them so that it can have coding gain. Let $\gamma_k^R$ be the SINR of the $k$-th trial at the receiver side including the HARQ coding gain, which can be expressed as the sum of increments of the HARQ coding gains:

$$\gamma_k^R = \sum_{j=1}^{k} \Delta_j,$$

where $\Delta_j$ is the incremental coding gain for the $j$-th received block and $\Delta_1 = \gamma_1$. For the coding gain $\Delta_j$ of $j = 2, \ldots, \bar{k}$, we adopt the work of Frederiksen et al. [11], [12] and modify it slightly as follows:

$$\Delta_j = \begin{cases} \gamma_j, & \text{if CC is adopted;} \\ g_{IR} \cdot \alpha_j \cdot \gamma_j, & \text{if IR is adopted,} \end{cases}$$

(1)

Note that extensive simulation results in [13] also supports their model.
where $g_{IR}$ is the coding gain of the incremental redundancy (IR) scheme, which is a function of the modulation, coding scheme and packet size [11]–[14]. Here, $\alpha_j$ is the ratio of the size of retransmitted block to that of the first block. By (1), we are assuming that the ratio of SINR increment per retransmission to original SINR, i.e., $\Delta_j/\gamma_j$, is proportional to $\alpha_j$. Since the energy of a redundancy block is $\alpha_j$ times that of the original packet, the HARQ combining model given in the above is fairly intuitive.

Let $P_k$ be the packet error rate after reception of $k$-th retransmitted packets. Then it can be expressed as a function of $\gamma_i^R$ and its data rate $r_k$:

$$P_k = P_e(r_k, \gamma_i^R) \text{ for } k = 1, \cdots, \bar{k}.$$  

where $r_k$ is the data rate of the selected MCS level at the $k$-th trial. Here, the loss rate function $P_e$ is decreasing in $\gamma_i^R$. Without loss of generality, we assume that $\lim_{\gamma \to \infty} P_e(r, \gamma) = 0$ for all $r$. If we assume that $P_i$ and $P_j$ are independent if $i \neq j$, then the probability $q_k$ that a successful transmission happens at the $k$-th trial is:

$$q_k := (1 - P_k) \prod_{j=1}^{k-1} P_j, \text{ for } k = 1, 2, \ldots, \bar{k}.$$  

Let $d_k$ denotes the occupied time-slots from the beginning of the first transmission until the completion of the $k$-th trial, then the average duration can be expressed as:

$$E[D] = \sum_{k=1}^{\bar{k}} d_k q_k + d_{\bar{k}} \sum_{k=\bar{k}}^{\infty} q_k.$$  

(2)

Let us define a period as the time from the first transmission until either a successful transmission or the $\bar{k}$-th trial. Then, in each period, either the original packet of size $B$ bit can be transmitted with probability of $\sum_{k=1}^{\bar{k}} q_k$ or nothing (0 bit) is transmitted, otherwise. Hence, the average throughput in bits at each period is

$$E[T] = B \sum_{k=1}^{\bar{k}} q_k.$$  

(3)

From the fundamental theorem of the renewal reward process [15], the throughput of the MS becomes the ratio

$$\frac{E[T]}{E[D]} = \frac{B \sum_{k=1}^{\bar{k}} q_k}{\sum_{k=1}^{\bar{k}} d_k q_k + d_{\bar{k}} \sum_{k=\bar{k}}^{\infty} q_k}.$$  

(4)

IV. ANALYTICAL THROUGHPUT MAXIMIZATION IN HARQ

For the ease of analysis, we assume that the channel varies very slow in the following subsection. With this assumption, we assume that $\gamma_k \approx \gamma_1$ for $k = 2, \ldots, \bar{k}$, which in turn implies that the transmission data rates are equalized as:

$$r_k = r_1 \text{ for } k = 2, \ldots, \bar{k}.$$  

However, for a given $\gamma$, $P_e(r, \gamma)$ is not a monotonic function of $r$ since it depends on the specific packet size used for each data rate $r$. 

‡
A. Slowly Varying Channel

As it takes $\tau_1 := \frac{B}{r_1}$ time-slots to transmit the first packet, at the $k$-th trial, it takes $\alpha_k \tau_1$ time slot from the assumption $r_k = r_1$. If we assume that the remaining time-slots between trials of the MS of interest are allocated to other MSs, the time-slots occupied till the completion of the $k$-th trial by the MS of interest $d_k$ is

$$d_k = \sum_{j=1}^{k} \tau_j = \tau_1 \sum_{j=1}^{k} \alpha_k,$$

with $\alpha_1 = 1$.

In addition, we assume that $\alpha_k = \alpha$ for all $k \geq 2$, otherwise maximizing (4) requires $(\bar{k} - 1)$-fold optimization. Furthermore, simulation results in Section VI (See Figs. 1 and 5.) substantiate the fact that the user throughput improvement is mostly achieved by small $k$. Now we can derive the following theorem (For readability, almost proofs in this paper are in Appendix.):

**Theorem 1.** Assume that $\alpha_k = \alpha$ for $k = 2, \ldots, \bar{k}$ and that the channel condition does not change until the $\bar{k}$-th retransmission. Then, the user throughput is simplified into

$$\frac{E[T]}{E[D]} = \frac{r_1(1 - \prod_{j=1}^{\bar{k}} P_j)}{1 + \alpha \sum_{k=1}^{\bar{k}-1} \prod_{j=1}^{k} P_j}. \quad (5)$$

In Theorem 1, we assumed that the wireless channel condition is constant or varies slowly. In other words, $\gamma_i$ is almost constant and does not change. This assumption is only valid for fixed or slowly moving mobiles and may not hold for fast moving mobiles. The assumption implies that the MCS (Modulation and Coding Scheme) level does not change for retransmitted packets, since the MCS level selection depends on $\gamma_i$ only. So, with the assumption of fixed $\gamma_i$, we can say that the MCS does not change for the retransmitted packets.

The throughput expression in (5) is applicable not only to the CC scheme but also to the IR scheme. The value of $\alpha$ is one and the incremental SINR value $g$ is one in the case of CC. Even though the expression (5) does not show $g$, $P_j$ is a function of the SINR of the $k$-th trial at the receiver side including the HARQ coding gain, $\gamma_i^{R_k}$, which includes the sum of the initial SINR $\gamma_1$ and the coding SINR.

In [10], a similar throughput expression is derived for the CC scheme. However, their equation is more complicated and is only applicable to the CC scheme. In contrast, our equation is simpler and is also applicable for the IR scheme.

**Lemma 1.** As $\alpha \to 0$, the expected throughput becomes

$$\lim_{\alpha \to 0} \frac{E[T]}{E[D]} = r_1(1 - P_{t_1}^{\bar{k}}). \quad (6)$$
Lemma 1 can be easily obtainable from equation (5). Note that (6) is still greater than the throughput of simple ARQ, which is \( r_1(1 - P_1) \) (See Section V-A). From a different standpoint, HARQ is approximated by simple ARQ with a packet error rate of \( P_1^\alpha \) for sufficiently small \( \alpha \).

**Lemma 2.** As \( k \to \infty \), the expected throughput becomes

\[
\lim_{k \to \infty} \frac{E[T]}{E[D]} = \frac{r_1}{1 + \alpha \sum_{k=1}^{\infty} \prod_{j=1}^{k} P_j}.
\]

(7)

As \( k \to \infty \), the average throughput \( E[T] \) in each period is now become constant \( B \), which is the result in Lemma 2. In a practical system, having \( k = \infty \) is impossible. However, we consider this case to gain some insights.

**B. Properties of the Throughput Optimization**

In general, upper protocol layers require that the packet error rate should be less than a certain threshold as we discussed in Section II. Thus we define the aggregate packet error rate as \( P_{\Pi} = \prod_{j=1}^{k} P_j \), which is the probability that the receiver fails in decoding after receiving the original packet and \( k - 1 \) redundancy blocks. The aggregate packet error rate should not be greater than the target aggregate packet error rate, which is denoted by \( P_T \). Therefore, for a given \( \gamma_1 \) and \( k \), we maximize (4) by varying \( \alpha \) and \( r_1 \), such that \( \prod_{j=0}^{\infty} P_j \leq P_T \). This constraint is ignored for the moment for mathematical tractability. Note that the throughput deterioration caused by this constraint is slight as shown in Section VI.

In the previous section, we derived throughput formula using the renewal reward process. If the formula is applied to a practical wireless system, \( \gamma_1 \) and \( k \) are predetermined so that only \( \alpha \) and \( r_1 \) can be varied. However, it is not clear what value of \( \alpha \) and \( r_1 \) would maximize (5) and (7). In fact, the maximization demands a complicated optimization with regard to \( \alpha \) and \( r_1 \) as (5) and (7) are nonconvex functions of the two variables. Let us consider the following cases: \( k \) is 1) limited and 2) unlimited.

While the size of retransmitted block is the same as that of the original packet in CC scheme, \( \alpha \) is different from 1 in IR scheme. The following lemma suggests the existence of the optimal \( \alpha \) that maximizes (5) in practical wireless packet data systems where \( k \) is finite.

**Lemma 3.** If equation (5) is continuous in \( \alpha \), there exists an optimal \( \alpha^* \) that maximizes (5) for a given \( k \).

In contrast, if we assume \( k = \infty \) which is impossible in real systems, we have the following lemma, which is rather surprising.

**Lemma 4.** Let \( \alpha^* \) be the optimal \( \alpha \) as a function of \( k \). As \( k \to \infty \), the maximum transmission rate \( mc_L \) is achievable with an arbitrary small \( \alpha^* \). That is to say,

\[
\sup_{\alpha \in (0, \infty), r_1 \in \{mc_1, \ldots, mc_L\}} \frac{E[T]}{E[D]} = mc_L.
\]
Proof: Since \(\lim_{\alpha \to 0} \lim_{k \to \infty} \frac{E[T]}{E[D]} = r_1\) and there is no restriction on the selection of the initial data rate \(r_1\) in (7), the above lemma holds as a matter of course. Alternatively, this result can be derived from (6).

This theorem implies that the maximum data rate \(m c_L\) is achievable for any \(\gamma_1\) by letting \(\alpha\) be arbitrarily small. Though the packet error rate \(P_1 = P_e(r_1, \gamma_1)\) increases as \(\gamma_1\) decreases in practical wireless systems, IR scheme overcomes the high packet error rate by transmitting an infinitesimal redundancy block for each retransmission request. Note that IR scheme should face with the problem of uncontrolled packet delays, which is sacrificed for achieving \(m c_L\). It is easy to see that the aggregate round-trip delay is proportional to \(1 + \sum_{k=1}^{\infty} \prod_{j=1}^{k} P_j\) where \(P_j\) approaches \(P_1\) as \(\alpha\) approaches 0. Hence, the aggregate round-trip delay asymptotically approaches \(\lim_{\alpha \to 0} 1 + \sum_{k=1}^{\infty} \prod_{j=1}^{k} P_j = \frac{1}{1-P_1}\) times a constant. In other words, the high user throughput in the above is achieved at the cost of uncontrollable delays.

V. DISCUSSION

A. Comparison of HARQ and ARQ Schemes

Let us consider two following cases assuming \(K \to \infty\). A simple ARQ sender retransmits the original packet and does not combine the retransmitted packets with the original packet. In this case, the supremum of the user throughput becomes \(\sup_{r_1} r_1 \cdot (1 - P_1)\) where the factor \(1 - P_1\) prevents simple ARQ from achieving data rates arbitrarily close to \(m c_L\). Readers are encouraged to derive the fact that the user throughput of simple ARQ does not depend on \(K\) due to its memoryless operation. Moreover, the user throughput of CC scheme is lower-bounded because \(1 + \sum_{k=1}^{\infty} \prod_{j=1}^{k} P_j \leq 1 + \sum_{k=1}^{\infty} P_1^k = \frac{1}{1-P_1}\). If the user throughput is regarded as a function of \(\alpha\), i.e., \(\frac{E[T(\alpha)]}{E[D(\alpha)]}\), the following relationship holds among HARQ/ARQ schemes:

\[
\frac{E[T(\alpha^*)]}{E[D(\alpha^*)]} \geq \frac{E[T(1)]}{E[D(1)]} \geq \sup_{r_1 \in \{mc_1,...,mc_L\}} r_1 \cdot (1 - P_1).
\]

B. On the Packet Overhead of IR Scheme

In the existing wireless packet data systems such as [16], even signals from the same cell are not orthogonal, but interfering with each other since those systems adopt code division multiple access (CDMA) for their downlink and uplink transmissions. Therefore, they prefer to make use of only time division multiplexing (TDM) to allocate air resources to different mobile stations. The minimum data size transmitted to a mobile station is quite bulky and the amount of data is not so flexible.

However, as prospective systems such as [1] and [2] adopt orthogonal frequency multiple access (OFDMA), the unit of data becomes much smaller than the existing systems. For example, the data
burst size used for downlink and uplink transmission in [2] is specified in DL-MAP (downlink map) and UL-MAP (uplink map) of each physical frame. The physical resource in a map to be occupied by a data burst forms a rectangle decided by two vertices where the first (second) vertex has two coordinates: 1) the position of the first (last) subcarrier to be occupied and 2) the position of the first (last) symbol to be occupied. Thus the evolving systems have the capability to finely adjust the redundancy block size of IR scheme. Moreover, as the maps are used for all physical frames regardless of HARQ schemes, there is no additional packet overhead required to let an IR receiver know the data burst size.

On the other hand, the ratio of the packet overhead to the payload containing systematic bits increases as the redundancy decreases. That is, an amount of systematic bits should be delivered to a receiver while the number of retransmission is kept as small as possible (condition 1). Furthermore, a resource block size in real systems should be an integer multiple of a system-specific amount of bits (condition 2). In the next section, we show that IR scheme performs much better than CC scheme does even if these two conditions are imposed on IR scheme.

VI. SIMULATION RESULTS

To make effective the condition imposed on the redundancy size in Section V-B, we assume that $\alpha$ is an integer multiple of $\alpha_\Delta$, i.e., $\alpha = n \cdot \alpha_\Delta$ for all integer $n \geq 1$. In this section, we use $\alpha_\Delta = 0.1$. As for the condition imposed on the number of retransmissions in Section V-B as well as Section II, we use only four $\bar{k}$ values, i.e., 2~5 to achieve small packet delays and to mitigate receiver buffers and packet overheads. Note that HARQ schemes are reduced to simple ARQ for $\bar{k}=1$.

We adopt the TDD mode of WirelessMAN-OFDMA 8.75 MHz channel basic PHY Profile in the IEEE 802.16e mobile broadband wireless access system for our simulation as it supports both CC and IR schemes [2] and its performance is watched with keen interest. There are 10 MCS levels whose data rates range from 650Kbps to $m_{CL} = 19.0$Mbps. The packet size used for each MCS level ranges from 60bytes to 480bytes. The packet error rates in additive white gaussian noise (AWGN) channel are obtained through extensive simulations based on the technical specification document [2].

Our last assumption is that $g_{IR} = 1$ all the time. It is apparent that the performance gain of IR scheme is more emphasized when $g_{IR}$ is greater than 1. However, it is not easy to optimize (5) since $g_{IR}$ weakly depends on the modulation, coding scheme and packet size in a complicated way [11]–[14]. Since IR scheme with $g_{IR} = 1$ always performs worse than that with $g_{IR} > 1$, simulation results in this work present lower bounds for situations in which $g_{IR} \geq 1$. Moreover, we found through simulations that the main benefit of IR scheme lies in its capability to finely tune the redundancy block size rather than in its intrinsic coding gain.

A detailed explanation of the frame structure of IEEE 802.16e OFDMA systems can be found in [17].
In simple ARQ scheme, denoted by ‘ARQ w/ $P_T = 0.01$’, an original packet is retransmitted for each unsuccessful reception at the noncombining decoder. The highest MCS level yielding the packet error rate ($P_1$) no more than $P_T = 0.01$ is used. In ‘CC w/o $P_T$’ and ‘CC w/ $P_T = 0.01$', two CC schemes are adopted with and without the condition on the aggregate packet error rates, $P_{\Pi} \leq P_T$, respectively. Two IR schemes, ‘IR w/o $P_T$’ and ‘IR w/ $P_T = 0.01$’, are defined in a similar way. In total, 17 schemes are simulated since 4 HARQ schemes is simulated for 4 different $k$ values, i.e., $1 + 4 \cdot 4 = 17$.

A. User Throughput Performance Under the Conditions on $k$ and $P_{\Pi}$

User throughputs of 17 schemes for various $\gamma_1$ and $\bar{k}$ are shown in Fig. 1. IR schemes significantly outperform CC schemes and simple ARQ scheme all the time. In particular, the superb performance of IR schemes is highlighted when $\gamma_1$ is near the MCS thresholds of simple ARQ. For example, if $\gamma_1$ is 14dB($=10^{1.4}$), two IR schemes for $\bar{k} = 4$ have the same throughput of 14.8Mbps, while two CC schemes for $\bar{k} = 4$ have 11.7Mbps and the simple ARQ has 11.4Mbps. Though CC and IR schemes have the same initial packet error rate ($P_1$) of 0.30, they have different values of $\alpha$, that are $1(=10 \cdot \alpha) \Delta$ and $0.1(=\alpha)$, respectively. It can be calculated that $\alpha = 0.1$ results in the residual packet error rate of $P_4 = 0.001$ and the aggregate packet error rate of $P_{\Pi} = \prod_{j=1}^{4} P_j = 3 \cdot 10^{-7}$, that are sufficiently small. Therefore, the superior performance of IR schemes at $\gamma_1=14$dB arises from the fact that the redundancy block size is sufficiently small whereas CC schemes waste the air resource by retransmitting the entire original packet. Similar phenomena occur throughout Fig. 1. It can be observed that $\alpha=1$ is far from optimal in most cases from a standpoint of user throughput.

For a given $\bar{k}$, user throughputs of CC schemes with and without the constraint $P_{\Pi} \leq P_T$ are indistinguishable in Fig. 1 since the effective SINR significantly increases per retransmission with $\alpha = 1$. Note also that the throughput difference between two IR schemes with and without the constraint $P_{\Pi} \leq P_T$ is also almost negligible. This signifies that the throughput-optimizing $\alpha$ in IR schemes seldom leads to high aggregate packet error rates. In other words, IR scheme effectively maximizes the user throughput while keeping the packet error rate under control.

Let us focus on the effect of $\bar{k}$. As shown in Fig. 1, the user throughputs of IR schemes are slightly different for various $\bar{k}$. However, observe that IR schemes even for $\bar{k}=2$ significantly outperforms CC schemes for $\bar{k}=5$, in user throughput. Since large $\bar{k}$ is undesirable and performance indices for $\bar{k}=3\sim5$ are nearly the same, $\bar{k}=3$ might be a reasonable choice for IR schemes, since both the high user throughputs and low packet delays are achievable with this choice.

The optimal redundancy size for different $\bar{k}$ and $\gamma_1$ is shown in Fig. 2 which reveals that $\alpha$ is disposed to decrease with $\bar{k}$ and $\gamma_1$. Note that this result coincides with the Lemma 4, which implies that optimal
α decreases with $\bar{k}$. Observe that the condition $P_{11} \leq P_T$ does not often result in the reduction of $\alpha$. Except for $\gamma_1 \leq 0$dB, the optimal $\alpha$ is less than or equal to 0.5, 0.4, 0.3 and 0.2, respectively for $\bar{k} = 2, 3, 4$ and 5. To simplify the selection of $\alpha$, a practical wireless packet data system might use these upper bounds (0.5, 0.4, 0.3 and 0.2) as its $\alpha$ for $\gamma_1 > 0$dB.

Optimal MCS thresholds are shown in Fig. 3. For a given $\gamma_1$, the MCS level of HARQ schemes is always equal to or larger than that of simple ARQ. Also observe that IR schemes adopt higher MCS level than CC schemes throughout the SINR. Note that HARQ schemes adopt very aggressive $\alpha$ values for low SINR in order to skip over low MCS levels such as 650Kbps and 1.3Mbps, as shown in Fig. 2. Since the ratios of channel rates in low MCS levels are very high, e.g., 1.3Mbps/650Kbps=2, HARQ schemes are inclined to skip over several MCS thresholds to adopt higher MCS levels. In other words, the gain obtained by adopting MCS levels is larger than the waste of time-slots incurred by adopting high $\alpha$. At low SINR, $\gamma_1 \leq 0$dB, MCS selection is rather irregular since both MCS levels and $\alpha$ are not continuous but discrete.

B. User Throughput Distribution in a Typical Cell

To see how the user throughput distribution looks like in a typical cellular system, let us consider the following SINR mapping: $\gamma_1 = \bar{p} \cdot g(d)/I$ where $\bar{p} = 40$dBm is the transmission power of the BS of interest and the thermal noise is neglected. The signal gain from the BS of interest to a user is defined as $g(d) \equiv \min \{1, \Gamma(d)\}$ where $\Gamma(d) = -130 - 37 \log_{10}(d)$ dB ($d$ is the distance from the BS $n$ to a user in kilometers). Note that this corresponds to a path loss exponent of 3.7. Let us assume that users are uniformly distributed over a circle cell with a radius of $\hat{d} = 2$km. The sum of other cell interference, $I$, is calculated as if there are six BSs that form a regular hexagon and are located $2 \cdot \hat{d}$ apart from the BS of interest.

Since the expected number of users who are located within the circular area, $[\hat{d}, \hat{d} + \Delta d]$ ($\Delta d$ is an infinitesimal), is proportional to $\hat{d}$ the SINR distribution of users appears like Fig. 4. So it is more likely that a random user is closer to cell boundaries than to the BS. Therefore, the advantage of IR is more emphasized when we consider the actual distribution of the SINR. Assuming that all user are allocated the same $\gamma_1$ and $\bar{k} = 3$, the cumulative distribution function of user throughputs can be obtained as shown in Fig. 5 where only representative schemes, i.e., ‘ARQ w/ $P_T = 0.01$’, ‘CC w/ $P_T = 0.01$’ and ‘IR w/ $P_T = 0.01$’, are considered. After dividing users into five equal-sized groups in order of throughput, we calculated the throughput increases of CC and IR schemes in each group and wrote them down in Fig. 5. Observe that IR scheme is superior to CC scheme throughout the SINR, while IR scheme is most effective for low SINR users. The problem of small data rates near cell boundaries, which is significantly
mitigated by IR scheme, is considered as one of difficult problems in next generation wireless packet data systems.

VII. CONCLUSION

Our contribution is three-fold: Firstly, we have presented a systematic and analytical framework that optimizes user throughputs in wireless packet data systems adopting general HARQ schemes by importing a moderately simplified model of HARQ combining gain. Moreover, the optimization framework provided in this paper is most suitable for identifying not only the optimal redundancy block size but also the optimal packet error rate and the the optimal mapping of signal to interference plus noise ratios into modulation and coding schemes.

Secondly, several mathematical characteristics of HARQ performance have been found. That is to say, the user throughput expressions derived in this paper are much simpler than those of previous work and pertain to be utilized in real wireless packet data systems. Moreover, we have proven that there exists an optimal redundancy block size when the number of retransmissions is limited and as small as possible redundancy block size is optimal when the number of retransmissions is unlimited.

Thirdly, to answer the question brought up in Section II, we have shown the excellent performance of IR schemes by simulating a real wireless packet data system. It is worth noticing that IR schemes significantly outperforms CC schemes in throughput while achieving both small packet error rates ($P_\Pi$) and low packet delays ($\bar{k}$). In particular, IR scheme is capable of mitigating the insufficient data rate problem of cell boundary users by skillfully interpolating MCS thresholds of simple ARQ. We can conclude that the strength of IR schemes mainly lies in its capability to adjust its redundancy block size adaptively rather than in its intrinsic coding gain.

APPENDIX: PROOFS

A. Proof of Lemma 1

The second term in the denominator of (5) can be upper-bounded as:

$$\alpha \sum_{k=1}^{\bar{k}-1} \prod_{j=1}^{k} P_j \leq \alpha \sum_{k=1}^{\bar{k}-1} P_1^k = \alpha \cdot \frac{P_1(1-P_1)^{\bar{k}-1}}{1-P_1}.$$ 

Therefore, it is easy to see that $\lim_{\alpha \to 0} \alpha \sum_{k=1}^{\bar{k}-1} \prod_{j=1}^{k} P_j = 0$ and hence the denominator in the limit becomes 1.

B. Proof of Lemma 2

From the assumption $\lim_{\gamma \to \infty} P_e(r, \gamma) = 0$ for all $r$, $\lim_{\bar{k} \to \infty} \prod_{j=1}^{\bar{k}} P_j = \lim_{\bar{k} \to \infty} \prod_{j=1}^{\bar{k}} P_e(r_j, \gamma_j^R) = 0$ for nonzero $\alpha$. So the numerator of (5) becomes $r_1$. 
C. Proof of Theorem 1

$E[D]$ can be simplified as follows.

$$
E[D] = \sum_{k=1}^{\bar{k}-1} d_k q_k + d_{\bar{k}} \sum_{k=\bar{k}}^{\infty} q_k = \sum_{k=1}^{\bar{k}-1} \tau_1 (1 + (k-1)\alpha) q_k + \tau_1 (1 + (\bar{k}-1)\alpha) \sum_{k=\bar{k}}^{\infty} q_k
$$

$$
= \sum_{k=1}^{\bar{k}-1} \tau_1 (1 + (k-1)\alpha) (1 - P_k) \prod_{j=1}^{k-1} P_j + \tau_1 (1 + (\bar{k}-1)\alpha) \sum_{k=\bar{k}}^{\infty} (1 - P_k) \prod_{j=1}^{k-1} P_j
$$

$$
= \sum_{k=1}^{\bar{k}-1} \tau_1 (1 + (k-1)\alpha) (1 - P_k) \prod_{j=1}^{k-1} P_j + \tau_1 (1 + (\bar{k}-1)\alpha) \sum_{k=\bar{k}}^{\infty} (1 - P_k) \prod_{j=1}^{k-1} P_j
$$

$$
\tau_1 \left( \sum_{k=1}^{\bar{k}-1} (1 + (k-1)\alpha) P_k - (1 + k\alpha) P_k + \alpha P_k \prod_{j=1}^{k} P_j \right) + (1 + (\bar{k}-1)\alpha) \sum_{k=\bar{k}}^{\infty} (1 - P_k) \prod_{j=1}^{k-1} P_j
$$

$$
= \tau_1 \left( 1 - (1 + (\bar{k}-1)\alpha) P_k + \alpha \sum_{k=1}^{\bar{k}-1} P_k + (1 + (\bar{k}-1)\alpha) \sum_{k=\bar{k}}^{\infty} (1 - P_k) \prod_{j=1}^{k-1} P_j \right)
$$

$$
= \tau_1 \left( 1 + \alpha \sum_{k=1}^{\bar{k}-1} P_k \right).
$$

It is easy to see that the expected number of successfully decoded bits is given by $E[T] = B \sum_{k=1}^{\bar{k}} q_k = B \sum_{k=1}^{\bar{k}} (1 - P_k) \prod_{j=1}^{k-1} P_j = B(1 - \prod_{j=1}^{\bar{k}} P_j)$. Thus, the user throughput in bits per seconds becomes:

$$
\frac{E[T]}{E[D]} = \frac{B(1 - \prod_{j=1}^{\bar{k}} P_j)}{\tau_1 \left( 1 + \alpha \sum_{k=1}^{\bar{k}-1} P_k \right)} = \frac{r_1 (1 - \prod_{j=1}^{\bar{k}} P_j)}{1 + \alpha \sum_{k=1}^{\bar{k}-1} P_k}.
$$

D. Proof of Lemma 3

Using the fact that $\lim_{\alpha \to 0} \frac{E[T(\alpha)]}{E[D(\alpha)]} = r_1 (1 - P_k)$ in Lemma 1, let us define $\frac{E[T(0)]}{E[D(0)]} := r_1 (1 - P_k)$. Since $\lim_{\alpha \to \infty} \frac{E[T(\alpha)]}{E[D(\alpha)]} = 0$, by the definition of limit, there exists an $\bar{\alpha}$ such that $\frac{E[T(\alpha)]}{E[D(\alpha)]} < \frac{E[T(0)]}{E[D(0)]}$ for all $\alpha \geq \bar{\alpha}$. Because the interval $[0, \bar{\alpha}]$ is a nonempty and compact set and $\frac{E[T(\alpha)]}{E[D(\alpha)]}$ is continuous on $[0, \bar{\alpha}]$, there should exist at least one optimal solution $\alpha^* \in [0, \bar{\alpha}]$ by Weierstrass’ Theorem [18] and the inequality $\frac{E[T(\alpha^*)]}{E[D(\alpha^*)]} \geq \frac{E[T(0)]}{E[D(0)]}$ is valid of course. By the definition of $\bar{\alpha}$, $\frac{E[T(\alpha^*)]}{E[D(\alpha^*)]}$ is also larger than $\frac{E[T(\alpha)]}{E[D(\alpha)]}$ for all $\alpha \geq \bar{\alpha}$ and becomes the maximum throughput in the interval $[0, \infty)$. 

REFERENCES

Fig. 1. Optimal user throughputs for various HARQ schemes.
Fig. 2. Optimal $\alpha$ for various $k$.

Fig. 3. Optimal MCS thresholds for various HARQ schemes.
Fig. 4. Probability density function of SINR.

Fig. 5. Cumulative distribution function of user throughputs for $k = 3$. 