Analysis of Network Coded HARQ for Multiple Unicast Flows

Peter Larsson†, Besma Smida†, Toshiaki Koike-Akino† and Vahid Tarokh†
†Ericsson Research, Isafjordsgatan 14E, 164 93 Kista, Sweden.
†School of Engineering and Applied Sciences, Harvard University, 33 Oxford Street, Cambridge, MA 02138, U.S.A.
Email: {plarsson, bsmida, koike, vahid}@seas.harvard.edu

Abstract—In this paper, we consider network coded (NCed) Hybrid-ARQ (HARQ) for multiple unicast flows. The main contribution of the paper is the derivation of throughput expression for NCed HARQ with arbitrary number of users i.i.d. channels. We apply the result to Rayleigh fading channels for incremental redundancy (IR) and chase combining (CC) based NCed HARQ. We verify the analytical approach with simulations. We observe substantial SNR improvements over regular ARQ with network coding as well as classical (H)ARQ. The SNR gains in the high and low throughput regimes are mainly due to the network coding and HARQ aspects, respectively. For low SNRs, NCed HARQ with IR surpasses the CC performance.

I. INTRODUCTION

With increasing number of radio devices, more communication use per device, and sprawling new high data rate services, there is a growing need for new transmission protocols enabling higher transmission rates. Moreover, the need for improved energy efficiency has become evident, not just due to the energy costs in operating cellular systems, or the desire of extending battery time of mobile handsets, but also due to the emergent need of cutting energy consumption in the society. Wireless cooperative communication techniques are unarguably a fertile ground for both reducing energy consumption and enhancing throughput. Notably, the recent idea of Network Coding (NC) [1] falls into the area of wireless cooperation. Yet, NC was originally considered for fixed wired and reliable links, but was soon found useful for unreliable links [2] and for noisy and lossy wireless channels [3].

The possibility of complementing NC with feedback for an unreliable wireless medium has not escaped the research community, as evident in a recent surge of work. However, we have found that already in ’84, Metzner proposed several NC-like ideas for multicast ARQ [4]. One idea was to send different linear combinations of packets until all users decoded a sufficient amount for extracting all packets. Moreover, even a random coding idea based on XORing was suggested. In ’93, Jolfai et al. also proposed the idea of using XORed packets for wireless multicast ARQ with feedback [5] and, presumably based on this, the idea was further explored in [6]. Since then, works like [7]-[9] have also studied NCed multicast ARQ.

The idea of using feedback based NC for unicast ARQ for multiple data flows has been considered in [10][11]. It was shown that in a block erasure channel, the throughput efficiency approached one with increasing number of users. While similar to [4]-[9], it turns out that the multiple unicast topology needs a different but more sophisticated coding strategy. An interesting aspect is that NCed unicast ARQ strive to project other user’s interfering data packets into a common subspace, orthogonal from the users own subspace. A similar philosophy is now the basis for Interference alignment [12]. NCed ARQ is also considered in [13]. A subsequent, but non-ARQ related, work on NC and unicast with feedback is [14].

For unreliable and noisy wireless channels, a variant of ARQ, Hybrid-ARQ (HARQ), has proven useful. HARQ exploits all received signal energy (rather than discarding some as in ARQ), which is beneficial to improve power and energy efficiency. Two basic HARQ approaches are Chase Combining (CC) [15][16] and Incremental Redundancy (IR) [17]. The former method (maximum ratio) combines retransmissions with noisy but identical signals. The latter one, more sophisticated, combines different codewords with identical information. A natural extension to NC and ARQ is to consider NCed HARQ. In [18], the authors analyzed and simulated NCed CC-HARQ for multicast. A NCed HARQ scheme for multiple unicast flows was studied through simulations in [19], where a 2 dB SNR gain over separate network and channel code was reported for block Rayleigh fading channels.

In this paper, we consider NC based CC- and IR-HARQ for multiple unicast flows. We introduce a coding design that enables both analytical tractability and good performance improvements. We derive a generic throughput expression for our NCed HARQ scheme applicable to different fading i.i.d. channels, modulation schemes, and both CC and IR-HARQ. Using mutual information (MI) [20], we derive and present analytical and simulated results for IR- and CC-HARQ in block Rayleigh fading. The main findings are: i) NCed-HARQ increases SNR sensitivity in the mid-to-high throughput range with increasing number of users, ii) for low SNRs, the gains over basic HARQ are modest, yet out-competing regular and NCed ARQ, and iii) IR-NCed HARQ has significantly better performance than NCed CC-HARQ in the low SNR regime.
II. NETWORK CODED HARQ

We consider a NCed HARQ system, illustrated by Fig. 1, with one sender and M receivers. Each user receives different data flows. We will first explain the coding/decoding procedure with reference to IR-HARQ and then we will indicate the differences for CC-HARQ.

A. Encoding and Scheduling

We have two type of packets, regular packets (RPs) and network coded packets (NCPs). An RP is transmitted in the form of a capacity achieving incremental redundancy code block (IR-block). The encoding of an RP into an IR-block uses different coding for each transmission. The encoding is assumed invertible i.e. the RP can be decoded from a single IR-block, and multiple IR-blocks can be code-combined for enhanced performance. An NCP have the same characteristics, but is composed of IR-blocks from several RPs. We further assume a network code with the property that a user, when knowing an NCP’s all but one packets and hence corresponding IR-blocks, can fully exploit the information contained in the unknown IR-block. For CC-HARQ, the same encoding is used for each CC-block, rather than generating new parity as for the IR-blocks.

In our studied NC and scheduling scheme, an RP is always sent individually first, before it may be part of an NCP. The RP may be a part of an NCP only if it has been decoded by one or more unintended users, but not its own. To form an NCP, we first select a group of m + 1 users where the users have previously decoded m RPs, i.e. one for every other user but not its own. Then, the m + 1 corresponding IR- or CC-block are composed into an NCP. With equal decoding probabilities, such NCP design is m + 1 times more throughput efficient than just sending an RP to one user, see [10].

B. Decoding

When receiving and decoding any own or other users RPs, each user uses code combining of IR-blocks for IR-HARQ, or maximum ratio combining of CC-blocks for CC-HARQ. Any correctly decoded own packet is forwarded to the next higher layer. Overheard, i.e. also decoded, packets are stored and later used for decoding of NCPs. At reception of an NCP designated for some users, a user has due to the scheduling method exact knowledge about all packets contained in the NCP apart from its own. Hence, all information carried by the NCP relates to one own’s data of interest only. So in combining the NCP with other code blocks derived from other RPs and NCPs, the relevant overheard decoded packets are exploited.

C. A Two-User Case

In Fig. 2, we illustrate this idea for two users. First, code blocks for an RP aimed for the first user are sent until any of the two users decodes it, which happen to be the second user here. Then, code blocks for the second user’s RP is sent, which is here decoded by the first user. As both users know each other’s RP, and can derive the associated code blocks, each user can extract their own code blocks in one or more subsequent and relevant NCPs and then combine them with the previously received but not yet decoded own code blocks.

D. Supplementary HARQ Operation

To facilitate the combining and decoding of NCPs, a user is informed (e.g. sent in a robustly encoded packet header) which packets are sent and the code block number. When a packet has been decoded the user return an ACK to the sender. The sender manages queues based on the received ACKs and uses the knowledge when scheduling and NC future transmissions. When packets are outdated or of no more use for NC, the sender may inform the users to discard overheard packets, or each user discards such packets after a timer expires.

E. Energy Consumption Aspects

From the simulation results in section IV, we note a promising energy saving potential. Yet, a rightful concern is the energy use for the overhearing of other user’s data, the additional signal processing needed, and the feedback. Regarding overhearing packets, as mainly the user on-time affects the energy use, this can be eased by exploiting OFDMA. A user then only overhears other user’s RPs when receiving its own RPs and NCPs. By always scheduling and NC for a set of users, a user’s on-time does not need to be increased. In fact, due to the improved throughput, a user’s on-time and corresponding energy use is reduced. Evidently, the extra ACK feedback requires some energy. Yet, this is negligible compared to the energy use of most cellular system basestations that consumes in order of kW’s [21] and user’s data transmission. Hence, if the emitted power can be reduced with just a few dB, then this still gives huge energy savings for the entire system as the main part of the power use relates to the power amplifier [21].

F. Complexity, Feedback, and Overhead Aspects

We recognize that complexity, feedback and overheard can be studied for NCed HARQ. In [13], the author raised such issues, though not scheduling complexity (SC) per se, for NCed ARQ. Due to space limits we will refrain from detailed studies on this here, and leave it for future research. Yet, in a preliminary assessment we find that the worst case mean SC per time is then 0, 6, 89 extra user pairing searches for M=2,6,10 users, which we think is acceptable. As will be seen, the largest relative gain increases are noted for few users.
The N Ced HARQ ACKs will also use extra uplink (UL) resources. Yet, as downlink (DL) often carries more traffic, and UL-DL spectrum are equal sized, it may be wise to use the UL to make the DL more throughput and energy efficient.

III. THROUGHPUT ANALYSIS

In this section, we derive a throughput expression for regular \((M = 1)\) and N Ced \((M \geq 2)\) Hybrid ARQ. For comparison, regular ARQ and N Ced ARQ are also considered. We assume identical packet sizes and data rates, non-sequence number limited HARQ, error free feedback, full buffers, ideal capacity achieving IR/CC-HARQ codes, and no signalling overhead (the payload is selected large relative the overhead).

A. Network Coded HARQ

Given \(M\) users, the rate \(R\), and the mean number of transmissions per packet \(S_M\), the throughput is defined as

\[ T_M = \frac{R}{S_M}. \]  

To ease the analysis, we divide the communication into two phases. In the first phase we send only RPs, and in the second phase packets can be sent in any order in each phase, but practically a packet may be resent until it is decoded, and then the next packet. We assume that a large number of packets are sent that allows for a statistical modeling of the relative amount of decoded packets. Let \(S'_M\) and \(S''_M\) be the mean number of transmissions per packet for the first and second phase respectively. \(S_M\) is then

\[ S_M = S'_M + S''_M. \]  

During the first phase, we ensure that each packet is at least successfully decoded by one user. The average number of transmissions is then

\[ S'_M = \sum_{k=1}^{\infty} k \prod_{j=0}^{k-1} (1 - Q_k^M) = \sum_{k=0}^{\infty} \prod_{j=0}^{k} Q_j^M, \]  

where \(Q_j\) is the probability that a user fails to decode a packet at its \(k\)th transmission given that this user exploit the information acquired from the \(k - 1\) earlier transmissions. We also set \(P_k = 1 - Q_k\) and \(Q_0 = 1\). We further define \(\tilde{Q}_k\) as the decoding failure probability after \(k\) successive transmission,

\[ \tilde{Q}_k = \prod_{j=0}^{k} Q_j. \]  

By inserting (4) in (3), we get

\[ S'_M = \sum_{k=0}^{\infty} \tilde{Q}_k^M. \]  

Note that \(S'_M\) is also known from [20] for the case of \(M = 1\).

It is however somewhat less easy to determine \(S''_M\). With a packet received by \(m\) users, \(m \in \{1, \ldots, M - 1\}\), in the first phase, a fraction \(m/M\) packets are decoded by the intended user. The fraction of packets for the second phase is then

\[ \rho(M, m) = \frac{M - m}{M}. \]  

As a starting point for the second phase, we need to consider the amount of information collected in the first phase by each user, when other user(s) decode its packets(s). The PDF for that a packet is decoded in the first phase at the \(j\)th transmission, \(j = \{1, \ldots, \infty\}\), by \(m\) users is

\[ P(M, m; j) = \binom{M}{m} Q_j^{M-m} P_j^m \prod_{u=0}^{j-1} Q_u^M. \]  

Given that a packet was decoded by \(m\) users at the \(j\)th transmission in the first phase, how many NCP transmissions are needed to deliver all incorrectly delivered packets to their intended users? This is given by the expectation of the required number of remaining transmissions

\[ S''_{rem}(j) = \sum_{k=1}^{\infty} \frac{k}{m + 1} P_{j+k} Q_j^{-1} \prod_{v=0}^{k-j-1} Q_v. \]  

where we account for the past failures in the first phase with the offset index \(j\), and where the factor \(m + 1\) is due to that NCPs is \(m + 1\) times more throughput efficient than an RP.

By summing (6), (7) and (8) over their indexes, we get the mean number of transmissions for the second phase as

\[ S''_M = \sum_{m=1}^{M-1} \rho(M, m) \sum_{j=1}^{\infty} P(M, m; j) S''_{rem}(j) \]
\[ = \sum_{m=1}^{M-1} \frac{M - m}{M} \times \prod_{j=1}^{M-j} Q_j (M - m) \prod_{j=1}^{M-j} P_j \prod_{j=1}^{M-j} Q_j^{-1} \prod_{v=0}^{k-j-1} Q_v \]
\[ = \frac{1}{M} \sum_{j=1}^{M-1} \prod_{u=0}^{j} Q_u (1 - Q_j)(1 - Q_j - MQ_j^{-1}) \]
\[ \times \prod_{v=0}^{k-j-1} Q_v. \]

where the last expression comes from summation over all \(m\). Now, using (4) in (9) gives us

\[ S''_M = \frac{1}{M} \sum_{j=1}^{\infty} \sum_{k=0}^{\infty} \left( \frac{\tilde{Q}_j - 1}{\tilde{Q}_j - 1 - M} \right) \tilde{Q}_j^{-1} \tilde{Q}_j. \]  

The exact throughput is subsequently found by inserting (10) and (5) in (2), and then in (1). For high \(M\) and SNR, accounting only for the first phase first transmission, (5) and (10) can be approximated with

\[ S'_M \approx 1, \quad S''_M \approx \sum_{k=1}^{\infty} \frac{\tilde{Q}_k}{M(1 - \tilde{Q}_1)}. \]

As \(P_1 = 1 - \tilde{Q}_1\) and \(T_1 = R/\sum_{k=0}^{\infty} \tilde{Q}_k\), this results in

\[ T_M \approx \frac{R}{1 + (R/T_1 - 1)/M P_1}. \]

This approximation turns out to be surprisingly exact at high throughput, and shows that \(\lim_{M \to \infty} T_M = R\) in Rayleigh fading.
B. Network Coded CC-HARQ in Rayleigh Fading Channel

In this section we determine $\tilde{Q}_k$ for CC-HARQ where we assume complex valued Gaussian distributed signals, AWGN, and block Rayleigh fading. With the fading modeled by the signal to noise ratio $x_u$, the accumulated mutual information (in nats) up to the $k^{th}$ transmission is

$$y^c_k = \ln(1 + \sum_{u=1}^{k} x_u). \quad (13)$$

The probability for a failed decoding at the $k^{th}$ transmission for a rate $R$ (in nats) message is in general

$$\tilde{Q}_k = \Pr(y_k \leq R) \quad (14)$$

that for CC-HARQ with (13) inserted can be rewritten to

$$\tilde{Q}^c_k = \Pr(z_k \leq \Theta), \quad (15)$$

where

$$z_k = \sum_{u=1}^{k} x_u \quad (16)$$

and $\Theta = e^R - 1$.

For a Rayleigh fading channel, each $x_u$ is i.i.d. exponentially distributed with the probability density function (PDF)

$$p_x = \lambda e^{-\lambda x}, \quad x \geq 0 \quad (17)$$

where $\lambda$ is the inverse mean SNR.

This makes $z_k$ gamma distributed with PDF

$$p_{z_k} = \frac{\lambda^k z_k^{k-1} e^{-\lambda z_k}}{(k-1)!}, \quad z_k \geq 0 \quad (18)$$

The probability that $z_k$ is less than the threshold $\Theta$ is

$$\Pr(z_k \leq \Theta) = \int_{0}^{\Theta} p_{z_k} dz_k = \gamma(k, \lambda\Theta), \quad (19)$$

where $\gamma(a,b)$ is the regularized lower incomplete Gamma function, $\forall a \in \mathbb{Z}^+, \gamma(a,b) = \int_{0}^{b} x^{a-1} e^{-x} \frac{dx}{(a-1)!}$, and $\gamma(0,b) = 1$.

Henceforth, we use the short hand notation $\gamma_k = \gamma(k, \lambda\Theta)$ when we now consider three interesting throughput cases: $M = 1$, $M = 2$ and a large value of $M$.

For $M = 1$, we have regular HARQ which does not use NC and hence no second phase. By swapping the order of the integration for $\gamma_k$ and the summation, we find

$$S_1 = \sum_{k=0}^{\infty} \gamma_k = 1 + \lambda\Theta. \quad (20)$$

Using (1), we get the single-user CC-HARQ throughput

$$T_1 = \frac{R}{1 + \lambda\Theta}. \quad (21)$$

For $M = 2$ and with (3), after some calculation, we have

$$S'_2 = \sum_{k=0}^{\infty} \gamma^2_k = 1 + \lambda\Theta - \lambda\Theta e^{-2\lambda\Theta} \left( I_0(2\lambda\Theta) + I_1(2\lambda\Theta) \right), \quad (22)$$

where $I_v$ is the modified Bessel function of the first kind of order $v$. Then, using (10) with $M = 2$, and with some further calculations, we get

$$S'_2 = \frac{1}{2} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} (\gamma_{j-1} - \gamma_j) \gamma_{j+k},$$

$$= \frac{\lambda\Theta e^{-2\lambda\Theta}}{2} \left( I_0(2\lambda\Theta) + I_1(2\lambda\Theta) \right), \quad (23)$$

With (22)(23) in (1), the two-user CC-HARQ throughput is

$$T_2 = \frac{R}{1 + \frac{\lambda\Theta}{M} e^{\lambda\Theta}}. \quad (24)$$

Note that, in contrast to when $M > 2$, this case is optimal as all overhead information is used for joint NC and HARQ. For large $M$ and reasonably high SNR, we use (12) to get

$$T_M \approx \frac{R}{1 + \frac{\lambda\Theta}{M} e^{\lambda\Theta}}. \quad (25)$$

C. Network Coded IR-HARQ in Rayleigh Fading Channel

For IR-HARQ, the accumulated mutual information is

$$y^c_k = \sum_{u=1}^{k} \ln(1 + x_u). \quad (26)$$

where

$$p_y = \lambda e^{\lambda+y - \lambda e^y}, \quad y \geq 0. \quad (27)$$

The decoding failure probability for the $k^{th}$ transmission can be calculated from

$$Q^c_k = \int_{0}^{R} p^k_y dy, \quad (28)$$

where $p^k_y$ is the $k$-fold convolution of $p_y$. Due to the integration intervals, it suffice to consider the $k$-fold convolution of a truncated, rather than infinite, version of $p_y$, where $p_y = 0$ for $y > R$. Also, the largest considered $k$ can be fairly small, say $k = 20$, as throughput is low when a large $k$ is required. As many values of $k$ is needed, the $k$-fold convolution may be calculated recursively $p^{k^*}_y = p^{k^*-1}_y * p_y$, or for reduced number of operations, De Pril’s method [22] may be used.

D. Network Coded ARQ

As a reference case, we include NCed (regular) ARQ. Then, the decoding failure probabilities $Q_1$ are constant. From (3), the mean number of transmissions per packet in phase one is

$$S'_M = \sum_{k=1}^{\infty} \prod_{j=0}^{k-1} Q_1 = \frac{1}{1 - Q_1^M}. \quad (29)$$
Correspondingly in the second phase, we have from (9)

\[
S_M' = \frac{1}{M} \sum_{j=1}^{\infty} \prod_{u=0}^{j-1} Q_1^M \frac{1 - Q_1^M - MQ_1^{M-1} P_1}{P_1} \\
\times \sum_{k=1}^{\infty} \prod_{v=j}^{k+j-1} Q_1, \\
= \frac{Q_1 (1 - Q_1^M - MP_1Q_1^{M-1})}{MP_1^2(1 - Q_1^M)},
\]

(30)

where \( P_1 = 1 - Q_1 \), and the last expression was found by calculating the two geometric series. In all, it gives us

\[
T_M = \frac{R(1 - Q_1^M)}{1 + M^{-1}P_1^{-2}Q_1(1 - Q_1^M - MP_1Q_1^{M-1})},
\]

(31)

which is the same result as in [10], but derived differently. For Rayleigh fading, we have \( P_1 = e^{-\lambda(2^{R}-1)} \).

IV. PERFORMANCE RESULTS AND DISCUSSION

This section presents both theoretical and simulated throughput results with respect to the mean SNR in block Rayleigh fading channels for network coded ARQ, CC-HARQ and IR-HARQ. When \( M \geq 2 \), the results corresponds to NC, whereas for curves in the figures with \( M = 1 \), i.e. a degenerated case of NC, corresponds to regular (H)ARQ.

Both a numerical evaluator and a simulator has been implemented in Matlab. The simulator uses the described two-phase method. We generate i.i.d. mutual information (MI) with relevant PDF for each link and transmission. Users accumulate corresponding MIs, and once the accumulated MI exceeds rate \( R \), a packet is considered decoded. We then count the number of transmissions required and determine the throughput.

Fig. 3 \((R = 1 \text{ b/Hz/s})\) and Fig. 4 \((R = 6 \text{ b/Hz/s})\) shows the throughput versus SNR for NCed ARQ, CC-HARQ and IR-HARQ with \( M = 1, 2, \ldots, 16 \) users. The continuous lines corresponds to theoretical results, and the associated markers ‘×’, ‘∗’ and ‘+’ signifies the simulation results. (The continuous line for CC-HARQ ends at low SNR as Matlab’s numerical precision of the incomplete gamma function is met.)

We observe the following: HARQ is significantly more SNR (and hence power/energy) efficient than ARQ at low SNR, NCed (H)ARQ improves throughput and SNR sensitivity with increasing number of users, NCed-(H)ARQ is significantly more SNR efficient relative to regular (H)ARQ for high SNR, NCed IR-HARQ becomes superior to NCed CC-HARQ first at higher rates \( R \), and all simulated results agree perfectly with the theoretical curves, which then validates our analysis.
Fig. 7. Exact (21), (24) and approximative (25) Regular and NC CC-HARQ throughput in \textit{i.i.d.} block Rayleigh fading vs. SNR.

It is also noted that for large \( M \) and large \( R \), NCed IR-HARQ develops a hump like characteristic. The hump at \( T=3 \) b/Hz/s in Fig. 4 corresponds to decoding success in the first phase for the second transmission. Subsequent humps, due to more retransmissions in phase one, exist but are successively less discernable. The humps for IR arise due to operation in the nonlinear region of \( \ln(1+\rho) \) at high SNR, where the resulting PDF becomes fairly compressed and hence localized.

Fig. 5 \((M = 1 \text{ user})\) and Fig. 6 \((M = 16 \text{ users})\) shows the throughput versus SNR for NCed ARQ, CC-HARQ and IR-HARQ with \( R = 1, 2, \ldots, 6 \) b/Hz/s. Again, the continuous lines corresponds to theoretical results and the associated markers ‘\( x \)’, ‘+’, and ‘\( * \)’ signifies simulation results.

In addition to the previous results, we observe the following: There are in general optimal rates \( R^* \) for each NCed (H)ARQ scheme as a function of mean SNR, with some reservation for NCed IR-HARQ. In the NCed IR-HARQ case, distinct optimal rates as a function of SNR develops first for higher SNR and higher number of users \( M \). This is so, since the hump like behaviour develops first at higher rates and number of users.

In Fig. 7, we evaluate and confirm the one- and two-user throughput expressions, (21) and (24), as well as the high-\( M \) throughput approximation (12) applied to CC-HARQ (25).

V. SUMMARY, CONCLUSIONS AND NEXT STEP

In this paper, we have presented a NCed based HARQ scheme. The core idea was to exploit collected information from earlier transmit attempts for both own, overheard and NCed packets. We proposed a scheduling and NC design that enabled analytical treatment and gave a general expression for the throughput expressed in terms of decoding failure probabilities. We then analyzed CC- and IR-based NCed HARQ in Rayleigh fading channels. Based on the throughput expression for NCed HARQ, we also derived an expression for NCed ARQ. A simulator was developed and its results for \textit{i.i.d.} block Rayleigh fading channels were shown with the analytical results, where we observed a perfect match. For the HARQ part of NCed HARQ, SNR gains were primarily visible in the low throughput range whereas the gains attributed more directly to NC were found in the high throughput range. We also observed that the throughput versus SNR curve steepened and shifted towards lower SNRs with increasing number of users that can be interpreted as form of multiuser diversity gain. We note that the observed SNR gains should be possible to translate into significant power/energy reductions. In the future we will look into ways how to more efficiently use all received/overheard information, consider delay aspects, non-equal mean SNRs, and common modulation formats.

ACKNOWLEDGMENT

The authors would like to thank their sponsors, the Marcus Wallenberg Foundation, the Swedish Research Foundation, the JSPS, and the SCAT.

REFERENCES