Hybrid ARQ optimizations
for wireless networks

Philippe Ciblat

Joint works with F. Bassi, P. Duhamel, A. Khreis,
N. Ksairi, A. Leduc, C. Le Martret, X. Leturc
Outline

- A very short introduction to Hybrid ARQ (HARQ)

- **HARQ improvement:**
  - *superposition coding*

- **HARQ parameters’ optimization:**
  - *adaptive modulation and coding scheme*

- **Resource allocation optimization** for HARQ based system:
  - *energy efficiency with Ricean channel*
Part 1 : Introduction to HARQ
Let \( S = [s_0, \cdots, s_{N-1}] \) be a packet composed by \( N \) uncoded symbols.

**Pros:**
- Adaptability to the real propagation states (noise, channel)
  - Robustness to no instantaneous Channel State Information at the Transmitter (CSIT)
  - Diversity if time-varying channel
  - High granularity with adapted Modulation and Coding Scheme (MCS) related to instantaneous channel behavior
- Cheap feedback link (one bit)

**Cons:**
- High latency, Buffer size
Remark

Retransmission does not contradict forward error coding (FEC)

Type-I HARQ: packet $\mathbf{S}$ is composed by coded symbols $s_n$

- first packet is more protected
- there is less retransmission
- transmission delay is reduced

- Efficiency is upper-bounded by the code rate

Drawbacks

- Each received packet is treated independently
- Mis-decoded packet is thrown in the trash
Type-II HARQ

Memory at RX side is considered ⇒ Type-II HARQ

Main examples:
- *Chase Combining* (CC)
- *Incremental Redundancy* (IR)
Examples: CC-HARQ and IR-HARQ

**CC**

\[
Y_1 = S_1 + N_1 \\
Y_2 = S_1 + N_2
\]

then detection on

\[
Y = \frac{(Y_1 + Y_2)}{2}
\]

SNR-Gain equal to 3dB

**IR**

\[
Y_1 = S_1(1) + N_1 \\
Y_2 = S_1(2) + N_2
\]

then detection on

\[
Y = [Y_1, Y_2]
\]

Coding gain
Delayed feedback management

Management for $T$:

- **Stop-and-Wait**

  
  
<table>
<thead>
<tr>
<th>$S_1(1)$</th>
<th>empty</th>
<th>empty</th>
<th>$S_1(2)$</th>
<th>empty</th>
<th>empty</th>
<th>$S_2(1)$</th>
</tr>
</thead>
</table>

  $T$ time-slots

  NACK

  ACK

- **Parallel Stop-and-Wait/Selective Repeat**

  
  
<table>
<thead>
<tr>
<th>$S_1(1)$</th>
<th>$S_2(1)$</th>
<th>$S_3(1)$</th>
<th>$S_1(2)$</th>
<th>$S_2(2)$</th>
<th>$S_3(2)$</th>
<th>$S_4(1)$</th>
</tr>
</thead>
</table>

  $T$ time-slots

  NACK  NACK  NACK  ACK

**Standard Assumption:**

- No error on feedback
- No delay ($T = 1$)
Performance metrics

- **Packet Error Rate (PER):**
  \[ \text{PER} = \text{Prob(} \text{message is not decoded} \text{)} \]

- **Efficiency (Throughput/Goodput/etc):**
  \[ \eta = \frac{\text{information bits received without error}}{\text{transmitted bits}} \]

- **(Mean) delay:**
  \[ d = \# \text{ transmitted packets when message is correctly received} \]

- **Jitter:**
  \[ \sigma_d = \text{delay standard deviation} \]

**Quality of Service (QoS)**

- Data: PER and efficiency
- Voice on IP: delay
- Video Streaming: efficiency and jitter
Closed-form expressions for metrics

\[
\text{PER} = 1 - \sum_{k=1}^{L} p(k)
\]

\[
\eta \propto \frac{\sum_{k=1}^{L} p(k)}{L(1 - \sum_{k=1}^{L} p(k)) + \sum_{k=1}^{L} kp(k)}
\]

\[
d = \frac{\sum_{k=1}^{L} kp(k)}{\sum_{k=1}^{L} p(k)}
\]

\[
\sigma_d = \sqrt{\frac{\sum_{k=1}^{L} k^2 p(k)}{\sum_{k=1}^{L} p(k)} - d^2}
\]

with \([\text{Leduc12}]\)

- \(p(k)\) probability to receive information packet in exactly \(k\) transmissions
- \(L\) maximum number of transmissions per message
Example: Type-I HARQ

- Let $\pi_0$ be the probability the message is not well decoded with one transmission. Then
  \[ p(k) = (1 - \pi_0)\pi_0^{k-1} \]

- Let a message be composed by $N$ BPSK uncoded symbols. Then, on Gaussian channel,
  \[ \pi_0 = 1 - \left(1 - Q\left(\sqrt{2\text{SNR}}\right)\right)^N \]

Results

\[
\begin{align*}
\text{PER} & = \pi_0^L \\
\eta & = 1 - \pi_0 \\
d & = L + \frac{1}{1 - \pi_0} - \frac{L}{1 - \pi_0^L} \\
\sigma_d^2 & = \frac{\pi_0 + \pi_0^{2L+1} - \pi_0^L(L^2 + \pi_0^2(1 + \pi_0)^2)}{(1 - \pi_0)^2(1 - \pi_0^L)^2} - \frac{2\pi_0^{L+1}(L^2 - 1)}{(1 - \pi_0)^2(1 - \pi_0^L)^2}
\end{align*}
\]
Part 2: HARQ optimizations

2.1 HARQ improvement
2.2 HARQ parameters’ optimization
2.3 Resource allocation optimization for HARQ based system
In practice, $T \neq 1$ ($T = 8$ in LTE)

**Idea for parallel Stop-and-Wait**

Send redundant packets **in advance** between pre-assigned time-slots and superpose them with packets related to other messages

- Similar idea when $T = 1$ [Shamai08, Assimi09, Szczecinski14]
- In SotA, with perfect CSIT or past CSIT (multi-bit feedback)
- Why could it work? non-orthogonal transmission with potential of Multiple Access Channel decoding

**Expected gains**

- Lower latency
- Higher reliability
2.1 - Proposed protocol

Let $p_k(\ell)$ be the $\ell$-th packet/chunk associated with the message $k$. We do a transmission with two layers:

- **Layer 1**: parallel Stop-and-Wait HARQ
- **Layer 2**: superposed redundant packets

$$p_k(\ell) \pm \sqrt{\alpha}p_k(\ell) + \sqrt{1 - \alpha}p_k'(\ell')$$

*without superposition*

*with superposition*

How do we choose the superposed redundant packets?

- Superpose packets of the most recent messages $\Rightarrow$ Low latency
- Superpose unsent redundant packets $\Rightarrow$ High reliability

---

Philippe Ciblat

Hybrid ARQ optimizations for wireless networks

12 / 35
2.1 - How to decode?

Received signal until time-slot 2

\[ y_1 = h(1)p_1(1) + w(1) \]
\[ y_2 = h(2)\sqrt{\alpha}p_2(1) + h(2)\sqrt{1-\alpha}p_1(2) + w(2) \]

Equivalent to a MIMO-MAC

Decoders:
- Multi-message based Decoder
- Single-message based Decoder

\[ R \]

Philippe Ciblat
2.1 - Numerical results: Throughput and MER

- **HARQ protocol**: IR-HARQ with $L = 3$, $R = 0.8$, best $\alpha$
- **Feedback delay**: $T = 3$ time-slots
- **Transmit energy**: $E_s$ (per symbol)

![Graphs showing throughput and MER vs. $E_s/N_0$ dB]

- Around 2dB-gain at moderate SNR
- 10% throughput gain at 0dB
- Diversity gain due to multi-layer transmission
2.1 - Numerical results: Latency

- More packets served with small delays (< 4 time-slots)
- but average delay close to each other
2.1 - Numerical results: Practical scheme

- IR-HARQ with $L = 3$
- RCPC whose successive rates are 0.8, 0.4, and 0.26.
- BPSK modulated symbols
- Decoding of message $k$:
  - Combining observations samples sharing the same packet $p_k(\ell)$
  - Calculating LLR for each observation sample
  - Computing Soft Viterbi’s algorithm
2.2 - Modulation and Coding scheme optimization

Main goal

Selecting the Modulation and Coding Scheme (MCS) per packet

- when IR-HARQ is used (packet=chunk)
- based on the available partially-outdated CSI

Why is it of interest?

- if BPSK: a few redundant bits sent but well protected
- if QAM: a lot of redundant bits sent but not well protected

LTE context

- Downlink from a base station (BS) to $K$ mobile users
- Transmission done by Resource Block (RB) $= Q$ channel uses
  - $B$ assigned RBs per frame and TX power per user constant during long duration, e.g., the so-called “semi-persistent scheduling” mode
  - MCS constant within 1 frame but adjustable frame by frame

$\rightarrow \text{MCS}_{k,t} \overset{\text{def}}{=} (m_k, t, R_k, t):$ the MCS during frame $t$
2.2 - IR-HARQ structure

- 1 HARQ round/transmission per frame
- When new transmission
  - Choose (by our algo.) a MCS with $2^{m_k,t}$ QAM and coding rate $R_{k,t}$
  - Thus $D_{k,t} = m_k,t R_{k,t} (QB)$ information bits to send
  - Apply a mother code of rate $R_0$ on these information bits
  - Then pick up $m_k,t (QB)$ coded bits to send
- When retransmission (after NACK): $D_{k,t}$ already fixed, and

$$R_{k,t} = \frac{D_{k,t}}{QB \sum_{j=t-\ell_k,t} m_{k,j}},$$

so choose $m_{k,t}$ only, such that, $R_{k,t} \geq R_0$. 

![IR-HARQ structure diagram](image-url)
2.2 - Channel model between BTS and user $k$

- **block fading and frequency selective:**
  - channel impulse response $h_{k,t} = [h_{k,t}(0), \ldots, h_{k,t}(M - 1)]^T$
    - constant on frame $t$
  - taps are independent and Rayleigh distributed

- **time correlated:**
  - $(h_{k,t})_t$ is a first-order Gauss-Markov process
    \[
    h_{k,t} = \alpha h_{k,t-1} + \sqrt{1 - \alpha^2} w_{k,t}(m), \quad t \geq 0.\]
  - $\alpha \in (0, 1)$ is the temporal fading coefficient from Jakes’ model

- **frequency response:** $N$-FFT of $h_{k,t}$
  \[
  H_{k,t} = [H_{k,t}(0), \ldots, H_{k,t}(N - 1)]^T
  \]
2.2 - Performance metrics: Block Error Rate (BLER)

\( \pi_{k,t} \): Block Error Rate (BLER) based on the so-far received frames

\[
\pi_{k,t} = 1 - d e^{-c_F \bar{\epsilon}_{k,t} - \cdots - c_1 \bar{\epsilon}_{k,t}}
\]

- \( F \in \mathbb{N}^* \): approximation order
- \( d, c_1, \ldots, c_F \): curve-fitting parameters
- \( \bar{\epsilon}_{k,t} \): average physical-layer Bit Error Rate (BER) associated with the hard-decision made on the so-far received coded bits [Vandendorpe09]

\[
\bar{\epsilon}_{k,t} = \frac{\sum_{j \in \text{HARQ rounds}} m_{k,t} \sum_{n \in \text{assigned channel uses}} 0.2 e^{-1.6 \frac{E_k |H_{k,j(n)}|^2}{(2^{m_{k,j}} - 1) N_0}}}{\# \text{of assigned channel uses} \times \sum_{j \in \text{HARQ rounds}} m_{k,j}}
\]
2.2 - Mathematical Goal

Our objectives

- For each frame, determine
  - the modulation scheme $m_{k,t}$
  - the coding rate $R_{k,t}$ (when new transmission)

- based on delayed CSI
  - $h_{k,t-1}$ or $\bar{c}_{k,t-1}$

- to maximize the average throughput

Appropriate tool: Markov Decision Process (MDP)
2.2 - MDP definition

**MDP framework**

- state space $S$: a set of states
- action space $A$: a set of *actions*
  - $A(s)$: admissible actions for state $s \in S$
- state transition distribution $Q(\cdot | s, a)$: $a \in A(s)$
- reward: a function $r : S \rightarrow \mathbb{R}$

**A deterministic policy**

A function $f : S \rightarrow A$

- such that $f(s) \in A(s), \forall s \in S$

**Goal**

Find out a policy maximizing an average reward
2.2 - Our MDP

- **State:**
  \[(\ell_{k,t}, h_{k,t-1}, \bar{c}_{k,t-1}, \# \text{of data bits in the codeword, } \# \text{of so-far received coded bits}) \in S\]
  with \(\ell_{k,t}\) corresponds to the number of previous transmissions but
  - \(\ell_{k,t} = 0\): a new transmission after an ACK
  - \(\ell_{k,t} = L\): a new transmission after \(L\) NACKs

- **Action:**
  - for new transmission: \((D_{k,t} = \# \text{of data bits}, m_{k,t})\) (29 values in LTE)
  - for retransmission: \(m_{k,t}\) (3 values in LTE)

- **Reward:**
  \[r(s) \overset{\text{def}}{=} \begin{cases} \frac{\# \text{of data bits in } s}{T_{\text{frame}}}, & \text{if } \ell = 0, \\ 0, & \text{otherwise,} \end{cases}\]
  and
  \[\eta_k^f(s_k,0) = \liminf_{t \to \infty} \frac{1}{t} \mathbb{E} \left[ \sum_{j=0}^{t-1} r(s_{k,j}) \right] \text{bits/sec}\]
Main result

Under mild assumptions, the problem is solvable, i.e., it exists an optimal policy \( f_k^* \) such that

\[
\eta_k^* = \eta_k^* \quad \text{with} \quad \eta_k^* \overset{\text{def}}{=} \sup_{f_k \in \mathcal{F}} \eta_k^f(s_k,0)
\]

Algorithm (Value Iteration)

\[
f_{k,t}(s) = \arg \max_{a \in A(s)} \left[ r(s) + \int_S v_{k,t-1}(y) Q(dy \mid s, a) \right]
\]

and

\[
v_{k,t}(s) = r(s) + \int_S v_{k,t-1}(y) Q(dy \mid s, f_{k,t}(s))
\]

with \( v_{k,0} = 0 \)
2.2 - Numerical results (LTE setup)

- In red, our proposed policy
- In blue, MCS leading to the smallest BLER $\overline{\epsilon}_{k,t}$ for the (outdated) $h_{k,t-1}$ (channel correlation not taken into account)
- In cyan, MCS leading to the smallest BER for the (outdated) $h_{k,t-1}$ (HARQ and channel correlation not taken into account)
- In black, random MCS

![Graph showing average throughput over subframe index with different schemes]

$\alpha = 0.969$ (v = 25 km/h)
2.3 - Statistical CSIT based resource allocation

Only channel statistics known at the transmitter

- fast-varying Rayleigh/Rice fading channel
- costly to report instantaneous channel realizations
- cheap to report statistics due to its coherence time

HARQ to handle unknown channel variation

Applications

- Mobile Ad Hoc networks (MANET)
- Cellular networks with high mobility
2.3 - Communication model

- **PHY layer within a cell**
  - OFDMA: no Inter-Symbol and Multi-User interferences
  - No multi-cell interference assumption

- **Statistical channel model (for the $k$-th link)**
  - Let $h_k(j, m)$ be the $m$-th filter tap at OFDMA symbol $j$
    - Independent but not identically distributed $\sim \mathcal{CN}(A_k \delta_m, 0, \varsigma^2_{k,m})$
  - Let $H_k(j, n)$ be the $n$-th Fourier component at OFDMA symbol $j$
    - Non-independent wrt $n$ but identically distributed $\sim \mathcal{CN}(A_k, \varsigma^2_k)$ with $\varsigma^2_k = \sum_m \varsigma^2_{k,m}$

**$k$-th link characterization**

- **Subcarriers are statistically equivalent**
  - $\gamma_k$: bandwidth proportion assigned to link $k$
  - $Q_k$: energy used by link $k$ in one OFDM symbol
    - Independent of subcarrier
    - $E_k = Q_k / \gamma_k$: energy of link $k$ in entire bandwidth

- **Rice fading channel**
2.3 - Resource allocation optimization problem

Energy-efficiency based problem

\[
\min_{\gamma, E} f\left(\{E_k(\gamma_k, E_k)\}_{k=1,\ldots,K}\right)
\]

s.t. \[ QoS_k(\gamma_k, E_k) \geq QoS_k^{(0)}, \quad \forall k \in \{1, \ldots, K\} \]
\[
\sum_{k=1}^{K} \gamma_k \leq 1
\]
\[
\gamma_k \geq 0, \quad E_k \geq 0, \quad \forall k \in \{1, \ldots, K\}
\]

with \( E_k \) = \# total amount of data correctly delivered by link \( k \)

\# total consumed energy on link \( k \)

the energy efficiency

Extensions:

- Cost functions: sum-goodput (MGO), sum-power (MPO)
- QoS constraints: MER, delay, goodput
2.3 - Why Energy Efficiency?

Example 1:

- **O1**: minimum power with goodput constraint (\( \geq 1\text{Mbits/s} \))
- **O2**: maximum goodput with power constraint (\( \leq 35\text{dBm} \))
- **O3**: maximum energy efficiency

Example 2: \( Q_r \) battery state (%), \( T_t \) time to transmit the messages (s), \( N_p \) number of transmitted messages, and goodput (Mbits/s)

<table>
<thead>
<tr>
<th>Full battery drain</th>
<th>( Q_r )</th>
<th>( T_t ) (s)</th>
<th>( N_p )</th>
<th>Goodput</th>
</tr>
</thead>
<tbody>
<tr>
<td>EE</td>
<td>0</td>
<td>8 327</td>
<td>( 2.8 \times 10^8 )</td>
<td>4.3</td>
</tr>
<tr>
<td>MGO</td>
<td>0</td>
<td>1 800</td>
<td>( 7 \times 10^7 )</td>
<td>5</td>
</tr>
<tr>
<td>MPO</td>
<td>0</td>
<td>12 180</td>
<td>( 9.5 \times 10^7 )</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>10^7 sent messages</th>
<th>( Q_r )</th>
<th>( T_t ) (s)</th>
<th>( N_p )</th>
<th>Goodput</th>
</tr>
</thead>
<tbody>
<tr>
<td>EE</td>
<td>96</td>
<td>297</td>
<td>( 10^7 )</td>
<td>4.3</td>
</tr>
<tr>
<td>MGO</td>
<td>85</td>
<td>256</td>
<td>( 10^7 )</td>
<td>5</td>
</tr>
<tr>
<td>MPO</td>
<td>89</td>
<td>1 280</td>
<td>( 10^7 )</td>
<td>1</td>
</tr>
</tbody>
</table>
2.3 - Practical optimization problem

Type-I HARQ with Rice channel and minimum goodput constraints

\[
\max_{\gamma, E} \sum_{k=1}^{K} m_k R_k \gamma_k (1 - q_k(G_k E_k)) \frac{\gamma_k}{\kappa_{1,k}} + \kappa_{2,k}\gamma_k E_k
\]

s.t. \( m_k R_k \gamma_k (1 - q_k(G_k E_k)) \geq \eta_k^{(0)}, \sum_{k=1}^{K} \gamma_k \leq 1, \gamma_k \geq 0, E_k \geq 0 \)

with

- \( G_k = |A_k|^2 + \varsigma_k^2 \)
- \( q_k \) probability that one frame in error

\[
q_k(G_k E_k) \approx a_k \left( b_k \sum_{\ell=1}^{4} c_\ell e^{-\frac{|A_k|^2 G_k E_k \theta \ell d_k}{1+\varsigma_k^2 G_k E_k \theta \ell d_k}} \right)^{\delta_k}
\]

Remark: Real MCS instead of information-theoretic metrics (like outage probability)
2.3 - How to solve it?

- \( f_k : x \mapsto 1 - q_k(G_kx) \) concave
- change of variables \((\gamma_k, E_k) \mapsto (\gamma_k, Q_k)\), then

\[
(\gamma_k, Q_k) \mapsto \gamma_k(1 - q_k(G_kQ_k/\gamma_k)) = \gamma_k f_k(Q_k/\gamma_k)
\]

is concave as perspective of \( f_k \)

Consequently, in \((\gamma_k, Q_k)\)
- Numerator: concave
- Denominator: convex (as linear)
- Constraints set: convex set

Results (fractional programming tool)

- Jong’s algorithm: solve at iteration \( i \) (with the above constraints)

\[
\max_{\gamma, Q} \sum_{k=1}^{K} u_k^{(i)} m_k R_k \gamma_k (1 - q_k(G_kQ_k/\gamma_k)) - v_k^{(i)} \kappa_{1,k} Q_k
\]

and update \( u_k^{(i)} \) and \( v_k^{(i)} \) according to well-defined equations

- KKT can be written in closed-form
2.3 - Numerical results

- $K = 10$ links, Bandwidth $W = 5$ MHz
- QPSK, convolutional code of rate 1/2
- Rician factor: 10 (dashed line), 0 (solid line)
Multi-layer HARQ: why does it work so well?

while multi-layer single-user communications w/o feedback is useless

\[ y = x_1 + x_2 + w \]

does not increase the capacity

\[ R = R_1 + R_2 < \log_2(1 + P_1 + P_2) = \log_2(1 + P) \]

with \( P \) the transmit power

- Design system with statistical CSI: time-varying Rician factor (joint work with French Thales company on MANET)

- Age of Information (AoI): relationship between HARQ and information refreshness
Our publications devoted to HARQ: a long story

**HARQ improvement:** superposition coding, IP level


**HARQ parameters’ optimization:** URLLC, MCS


**Resource allocation optimization** for HARQ:


