

CODE RATE OPTIMIZATION UNDER FINITE BLOCK-LENGTH USING HYBRID ARQ FOR LOW LATENCY COMMUNICATION

by

Muhammad Zulqarnain

2015-06-0020

A thesis submitted in partial fulfilment of the requirements
for the Degree of

MASTER OF SCIENCE
in
Electrical Engineering

Supervisor: Momin Ayub Uppal



Department of Electrical Engineering
Syed Babar Ali School of Science and Engineering
Lahore University of Management Sciences

June 2017



To Whom It May Concern

The document entitled “**Code Rate Optimization under Finite Block-Length Using Hybrid ARQ for Low Latency Communication**” submitted by **Muhammad Zulqarnain**; Roll # 2015-06-0020 of Department of Electrical Engineering; Syed Babar Ali School of Science and Engineering, was checked through Turnitin on June 8, 2017 to determine if it is plagiarism free or otherwise. After excluding quoted material, bibliographies and small matches (up to 25 words) the originality report indicates that similarity index is **02 (Two)** percent that:

- a) Meets the Higher Education Commission of Pakistan required standard (up to 19 %). ✓
- b) Does not meet the Higher Education Commission of Pakistan required standard (up to 19 %).

08/06/2017

Waris Ali Arslan

Focal Person for Plagiarism Check

Deputy Manager

Gad & Birgit Rausing Library

waris.arslan@lums.edu.pk

Abstract

The envisioned applications of 5G-networks involve more robust and efficient machine-type communication (C-MTC) where seamless control requires ultra high reliability and extremely low latencies on the order of 10^{-9} and 1 ms, respectively. Achieving these stringent requirements over a wireless network is a challenging task. As a step towards developing low-latency network, this thesis investigates the effect of coding rate on latency and reliability under a finite block-length regime. We study the tradeoffs that channel code rate introduce in three specific cases: (a) Automatic Repeat Requests (ARQ), (b) Hybrid ARQ-Chase Combining (HARQ-CC), and (c) Hybrid ARQ-Incremental Redundancy (HARQ-INR). Our results indicate that for a given reliability constraint, HARQ-INR always performs better than HARQ-CC. Moreover, the results also show that HARQ-CC performs better than ARQ, especially in the low-SNR regime.

Acknowledgements

At the very outset and with the bottom of my heart I would like to express all my devotion to Almighty Allah, most merciful who enabled me to complete my thesis for the fulfillment of my degree. Secondly, I would like to thank my advisor Dr. Momin Ayub Uppal for his valuable advice and guidance and exceptional analytical skills which enabled me with the motivation of learning and research. I would also like to thank my fellow colleagues and my family. In the end, I would like to thank Faculty of Electrical Engineering at LUMS, as they have put high energy in me with the spirit of learning and innovating.

Contents

Abstract	i
Acknowledgements	ii
List of Figures	v
List of Tables	vi
List of Abbreviations	vii
1 Introduction	1
1.1 Low Latency: A 5G Dream	1
1.1.1 Challenges Associated	2
1.2 The Tactile Internet and Applications	3
1.2.1 Health Care	4
1.2.2 Smart Grids	4
1.2.3 Vehicle Mobility	4
1.3 Thesis Organization and Contributions	5
2 Communication under finite block-length regime	7
2.1 Maximal code rate R under a finite block-length	7
2.2 Communication Model	9
2.3 Latency	11

2.4	Hybrid ARQ	11
2.4.1	Chase Combining	12
2.4.2	Incremental Redundancy	14
3	Literature Review	16
3.1	Co-operative Strategies	16
3.2	Air Interface designs	18
4	Coding Rate and Latency	19
4.1	Latency Formulation under ARQ	19
4.1.1	Average no. of transmissions under ARQ	19
4.1.2	Optimization problem for ARQ	24
4.1.3	Bounds on R for minimizing Latency in ARQ	25
4.2	Latency Formulation under Hybrid ARQ	28
4.2.1	Average no. of channel uses in Hybrid ARQ	28
4.2.2	Optimization problem for Chase Combining	29
4.2.3	Solution to N_{max} under Chase Combining	31
4.2.4	Optimization problem for Incremental Redundancy	31
5	Simulation Results	33
5.0.1	Performance of ARQ at different SNR	33
5.0.2	SNR vs Latency comparison under Chase Combining	35
5.0.3	Performance of Incremental Redundancy at different SNR	36
5.0.4	Performance analysis of Hybrid ARQ vs ARQ	37
6	Conclusion and Future Work	39
	Bibliography	43

List of Figures

1-1	Fundamental Tradeoff between Latency, Reliability and Throughput [1]	2
2-1	Communication System Model	8
2-2	Hybrid ARQ Chase Combine Block Diagram	12
2-3	Hybrid ARQ Incremental Redundancy Block Diagram	14
4-1	Visual Representation of the cases of N_p	21
5-1	Performance of ARQ at different SNR	34
5-2	Performance of Chase Combining at different SNR	34
5-3	N_{max} in Chase Combining at different SNR	35
5-4	Performance of Incremental Redundancy at different SNR	36
5-5	Optimal Latency vs SNR for all Schemes	37
5-6	Spectral Efficiency vs SNR for all Schemes	38

List of Tables

5.1	Optimal Code Rate for different SNR in ARQ	33
5.2	Optimal Latency & Code Rate for different SNR in Chase Combining	35

List of Abbreviations

3G	3 rd Generation Mobile Networks
3GPP	3 rd Generation Partnership Project
4G	4 th Generation Mobile Networks
5G	5 th Generation Mobile Networks
ACK	Acknowledge
AF	Amplify and Forward
ARQ	Automatic Repeat Requests
AWGN	Additive White Gaussian Noise
C-MTC	Critical Machine Type Communication
DF	Decode and Forward
HARQ	Hybrid Automatic Repeat Requests
HARQ-CC	...	Hybrid Automatic Repeat Requests – Chase Combining
HARQ-INR	..	Hybrid Automatic Repeat Requests – Incremental Redundancy
IEEE	Institute of Electrical & Electronics Engineers
IoT	Internet of Things

LTE	Long Term Evolution
M-MTC	Massive Machine Type Communication
MAC	Media Access Control Layer
MRC	Maximum Ratio Combining
MTC	Machine Type Communication
NAK	Not Acknowledge
OccypyCOW		Optimizing Cooperative Communication for Ultra-reliable Proto- cols Yoking Control Onto Wireless
OFDM	Orthogonal Frequency Domain Multiple Access
PHY	Physical layer
SERCOS	Serial Real-time Communication System
SNR	Signal to Noise Ratio
TDMA	Time Division Multiple Access
XOR-CoW	...	Exclusive OR Control Onto Wireless

Chapter 1

Introduction

In the near future, massive numbers of ubiquitously distributed, mobile embedded systems and access devices that will communicate with each other and transfer large amounts of data at high speed with high availability will become a reality. These devices need high-speed communication with ultra high reliability and thus requires ultra low latencies where achieving them is a challenging task. This thesis strives to overcome this problem by optimizing code rate in order to meet these demands under some performance constraints in a finite block-length regime. This chapter introduces potential applications on low latency communication, associated challenges and the organization of this thesis with its contributions.

1.1 Low Latency: A 5G Dream

In previous generations of communication (3G & 4G) the primary focus was to increase throughput with high reliability, 5G envisions much more than that by connecting every thing with each other at high speed and faster response time, thus enabling Tactile internet (a dream in which one can move an actuator at a remote location with touch of a finger). For this case people have started working on Internet of Things from two aspects: One is Massive machine type communication (M-MTC) and other is

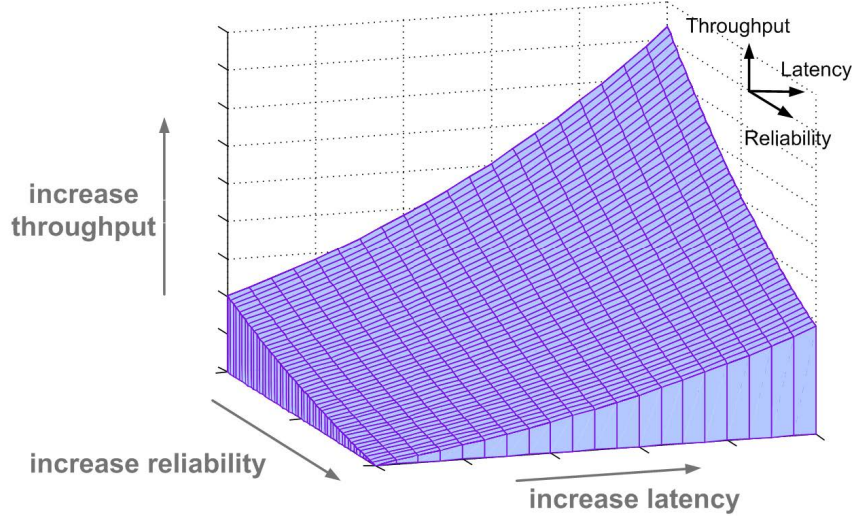


Figure 1-1: Fundamental Tradeoff between Latency, Reliability and Throughput [1]

Critical machine type communication (C-MTC). M-MTC focuses on connecting large number of devices such as sensors, actuators at the same time with high requirements of coverage, energy and bandwidth efficiency. Whereas, C-MTC focuses on critical communication between these devices which requires very low latency on the orders of less than 1 ms and high reliability of the order of 10^{-9} . Current control systems and automation industry relies highly on cable communication and follows standards like SERCOS-III, Profibus etc. As wired communication have major drawbacks like installation, maintenance of cables, protection and insulation, wireless provides high mobility and flexibility. But implementing it for C-MTC is a challenging task due to low latency requirements and high reliability which current wireless standards cannot deliver.

1.1.1 Challenges Associated

Though achieving the standards of SERCOS-III in a wireless medium is challenging but it is doable. Achieving low latencies, high reliability and high throughput simultaneously is harder as there are trade-off associated. Figure 1-1 shows these trade-off

in achieving high throughput with high reliability and low latency. One can achieve high throughput with high reliability by bearing the cost in increased latency, and this was the motivation behind 4G. Focus of C-MTC is to optimize these trade-offs for the required target application. Where reliability is defined as the probability of error in decoding a packet, latency as a delay in terms of average no. of channel uses per bit and throughput as the maximum rate at which communication is possible, which for a given channel conditions was derived by Claude Shannon in 1948 [2] for an infinite block-length in the limit of infinite no. of channel uses.

1.2 The Tactile Internet and Applications

Low latency and high reliability brings a real-time sensation. As a human being we want real-time interaction with things around us like driving cars, video-conferences, making phone calls etc. Engineers define a service as a real-time when the communication response time is faster than the time constants of an application. If the communication response time is more than the application processing and response time than we face cyber sickness. One could ask: How minimum should a communication delay be ?. The answer lies in the target application as different kinds of human sensations are sensitive to different types of applications. If a voice telephone have a delay up to 100 ms this will go un-noticeable as our ears and brain are robust to this much delay. But if we talk about an application which requires muscle movement e.g. reacting to another person touch, response time reduces to about 1 second. So, for an tactile application which requires ultra-low latency and high reliability like virtual reality and augmented reality a latency requirement of 10 ms is satisfactory, in order to synchronize frames and audio between eye and ear to avoid simulator sickness [3]. As many applications can afford latencies up to 10 ms and still provide real-time sensation, there are many which require round-trip time less than 1 ms and

they are currently relying on wired standards. Few such potential applications of tactile internet are discussed:

1.2.1 Health Care

Health care is a multi-diverse discipline requiring low latency demands from remote health-care to tactile control of exoskeleton. One such experiment was performed by University of California Berkeley's in which they have used exoskeleton (a device that uses body limbs to generate force greater than muscles) to give the ability to disable persons to walk again. Exoskeletons use actuators and sensors which are controlled by a controller over a wireless channel. These sensors and actuators require ultra-low latency and high reliability for the person to walk properly. This low latency will open gates to many new experiments in the field of Health Care.

1.2.2 Smart Grids

Smart Grids is a network of electricity which uses communication technology to control the power flow in a grid station. This again requires low latency and high availability of the network to have a maximal efficiency between the supply and demand of power.

1.2.3 Vehicle Mobility

Round trip of latency of 1 ms will not only open new gates in the field of Health care, industrial automation and other but in the field of vehicle mobility as well. e.g. In a bad weather scenario if a driver feels unsafe to drive then vehicle can be remotely controlled by a call center which again requires low latencies to have a real time feedback loop to avoid accidents. Not only this but whole metro bus system, railways can be made automated by having reliable remote control over the vehicles.

This chapter just highlights few examples of C-MTC which can only become a reality if we achieve low latencies and high reliability demands of tactile internet.

1.3 Thesis Organization and Contributions

Our objective in this thesis is to find the optimal solution which allows one to achieve low latency with tolerable probability of error. Though latencies are caused by multiple factors like processing delays of a digital hardware, distance between nodes, traffic of network, our focus in this thesis is to minimize latencies introduced by number of channel uses per bit by fixing block-length. We achieve this task by observing the effect of code rate on number of channel uses per bit in a finite block-length regime for three cases:

- Automatic Repeat Requests (ARQ)
- Hybrid ARQ — Chase Combining (HARQ-CC)
- Hybrid ARQ — Incremental Redundancy (HARQ-INR)

Where in traditional ARQ if a receiver fails to decode a packet it discards it and the same packet is re-transmitted by the transmitter whereas in HARQ cases the failed packets are saved in a buffer and are used to combine with new received packet in way that SNR is maximized (Chase Combining) or block-length is increased (Incremental Redundancy). We develop optimization problems for these three cases by constraining under network reliability which for tactile internet is 10^{-9} and finding the optimal code rate for given channel conditions. In chapter 2, we briefly discuss the effects of finite block-length on channel capacity by using the results derived by [4] and a brief introduction on Hybrid ARQ. Chapter 3 discuss some approaches which are discussed in literature and chapter 4 derives mathematical formulation for average number of channel uses under ARQ and Hybrid ARQ cases and forms an optimization problem

for the aforementioned cases. Chapter 5 provides simulated results and finally chapter 6 concludes this study.

Chapter 2

Communication under finite block-length regime

This chapter introduces fundamentals of a communication system while defining latency as a function of average number of channel uses per bit. We also introduce some fundamental results from [4] for a communication under finite block-length with some tolerable probability of error. Furthermore, a brief overview of Hybrid ARQ schemes: Chase Combining (HARQ-CC) and Incremental Redundancy (HARQ-INR) is provided.

2.1 Maximal code rate R under a finite block-length

According to the classical result [2], Shannon's capacity formula says that for a given SNR one can transmit a certain number of bits per channel use. The Capacity formula is given by:

$$C = \log_2(1 + \text{SNR}) \tag{2.1}$$

The result of (2.1) says that a certain person can transmit up to 1 bit per channel use for a given SNR of 0 dB with vanishing probability of error under the limit of

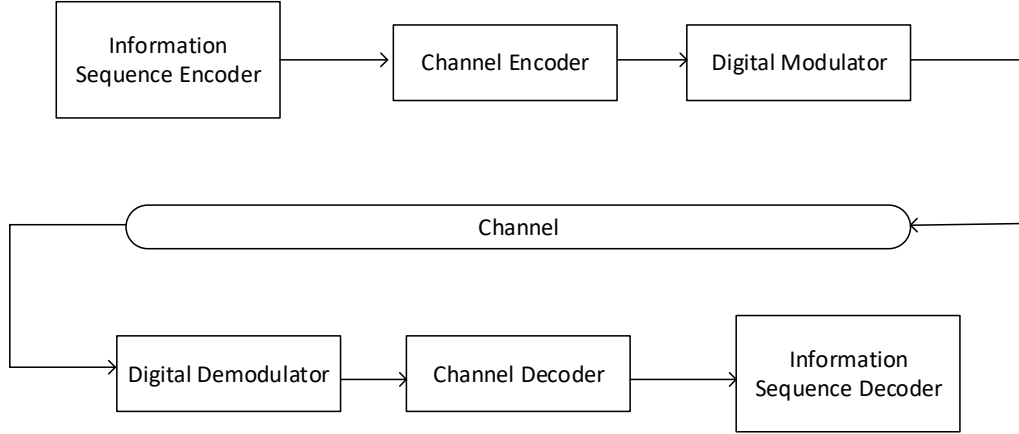


Figure 2-1: Communication System Model

an infinite number of channel uses, thus requiring infinite block-length. Whereas, in practice we can't have an infinite block-length as it is defined as a duration of a communication which results in infinite latency. And in C-MTC we require block-length to be short of the orders of ~ 100 channel uses [5], [6]. So, we need to use finite block-length which means that we suffer plenty in the form of maximal rate achievable. Where expression for maximal rate under a finite block-length regime was approximated by [4] and is given as:

$$R = C - \sqrt{\frac{V}{n}} Q^{-1}(\varepsilon) \quad (2.2)$$

where C is the channel capacity, V is the channel dispersion, n is the block-length and ε is the probability of error. Following section provides brief introduction of these terms.

2.2 Communication Model

A complete communication system as shown in Figure 2-1 is responsible for reliable delivery of message in a fastest way possible. A basic Communication Model comprises of the following blocks:

- **Information Source Encoder:** This block converts the digital or analog message to binary format and outputs binary data.
- **Channel Encoder:** This block adds redundancy to the message by adding n coding information to k information bits. The rate at which this block adds redundancy is called the code rate $R = \frac{k}{n}$.
- **Digital Modulator:** Before transmitting message over a channel the message must be modulated so it can travel through channel required frequency support.
- **Channel:** This is modelled as the noise that is added to message when transmitted. In all communication electronics AWGN noise is always present.
- **Digital Demodulator:** The purpose of this block is to map the received message to the message set based on probability of error and modulation type used.
- **Channel Decoder:** This block decodes the message based on coding used in channel encoder.
- **Source Decoder:** Message is converted back into the Analog or digital whichever is desired.

We would like to describe some basic terms that will be used throughout later in this context.

Channel Reliability p^*

Reliability is the measure of message being successfully received and decoded. It is the probability of error that occurs when a message is not successfully received or fails to decode by channel decoder. This is the performance metric of our thesis and for low latency networks it is set as $p^* = 10^{-9}$.

Channel Capacity C

Channel Capacity is the maximum data rate at which reliable communication over a certain channel is possible [2]. Under Shannon's limit this is given as:

$$C = \log(1 + P) \quad (2.3)$$

where P is the SNR (signal to noise ratio).

Channel Dispersion V

For a fixed capacity the probabilistic variation in the channel relative to its deterministic equivalent is known as channel dispersion [7]. The expression for channel dispersion for an AWGN channel under a finite block-lengths is given as:

$$V = \frac{P}{2} \frac{P+2}{(P+1)^2} \log^2 e \quad (2.4)$$

The error function ε for a maximum rate R

According to [8] the maximum coding rate R achievable for a given block-length n and error probability ε is given by (2.2). Where after re-arranging 2.2 we can have an error expression for a given block-length and a maximum code rate as:

$$\varepsilon = Q \left((C - R) \sqrt{\frac{n}{V}} \right) \quad (2.5)$$

where Q is complementary Gaussian cumulative distribution function.

2.3 Latency

Latency is defined as the average no. of channel uses which is given by:

$$\text{Latency} = \mathcal{T} \times n \quad (2.6)$$

where \mathcal{T} is the average number of packets in a transmission and n is the block-length which is considered as a duration of a communication. Normalizing (2.6) over information bits k we have:

$$\text{Latency} = \mathcal{T} \times \frac{n}{k} \quad (2.7)$$

Hence, (2.7) defines latency as the average no. of channel uses per bit. One can find latency in seconds by using (2.6) and normalizing with symbol duration T . In chapter 4 we will provide analytical expressions for \mathcal{T} for ARQ and HARQ and we will use (2.7) to form optimization problem under fix block-length with the assumption of known SNR.

2.4 Hybrid ARQ

In traditional ARQ a packet is transmitted by transmitter and if the receiver fails to decode that packet it is re-transmitted by transmitter after receiving NAK by the receiver. In Hybrid automatic repeat request (HARQ), if a packet fails the receiver saves the copy of that packet in a buffer and sends NAK to transmitter, upon re-transmitting the packet the receiver combines the new received packet with previous one on the basis of SNR maximization (Chase Combining) and lower code-rate (Incremental Redundancy). Both of these schemes are fall under the category of Hybrid ARQ soft combining [9]. Performance comparison of HARQ is studied in various

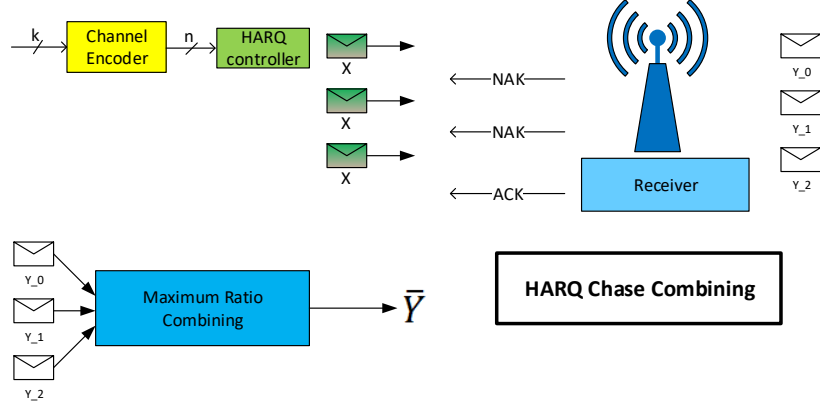


Figure 2-2: Hybrid ARQ Chase Combine Block Diagram

literatures e.g. [10], [11], [12] where it is shown that how Hybrid ARQ can be used to combat packet failures in fading wireless channel. All of these papers considers performance of Hybrid ARQ under the assumption of an infinite block-length. Two schemes incorporating HARQ with different variants are described below:

2.4.1 Chase Combining

Chase combining uses Maximum ratio combining to combine the subsequent packets in a buffer in order to maximize the SNR. According to system model shown in Figure 2-2 MRC is used at the receiver to re-combine packets in a buffer such that noise is reduced and SNR is maximized. Hence, Chase combine tends to maximize SNR in noisy channel. In this scheme identical copies of the same packet are re-transmitted until the packet is fully decoded by the receiver whereas in partial chase combining only subset of bits of the original transmission are re-transmitted. These bits can be error correcting codes or part of payload which needs correction. A Mechanism of Chase Combining is described below:

- Let the message be $X[i]$ where i is the symbol index.
- Message is re-transmitted upon receiving NAK so $X[i]$ is fixed

- Channel model is given as:

$$Y[i] = hX[i] + Z[i] \quad (2.8)$$

where h is the fading co-efficient and $Z[i]$ is the noise.

- We will have the following observation at the receiver:

$$Y_j[0], Y_j[1], \dots, Y_j[n-1] \quad (2.9)$$

Suppose the receiver fails to decode the j^{th} trial then a NAK is send back to Transmitter which upon receiving the transmitter re-sends the same message and now the new trial is:

$$Y_{j+1}[0], Y_{j+1}[1], \dots, Y_{j+1}[n-1] \quad (2.10)$$

Now the receiver combine these two trials and attempts to decode the message using MRC. MRC will be simply the average of two trials. If the receiver fails to decode using these two trials then the next trial $j+2$ is requested to decode the message. At the end of j^{th} trial average value after the receiver is:

$$\begin{aligned} \bar{Y}_j[i] &= \frac{1}{j+1} \sum_{k=0}^j Y_k[i] \\ &= \frac{1}{j+1} \sum_{k=0}^j (hX[i] + Z_k[i]) \\ &= \frac{1}{j+1} \left[(j+1)hX[i] + \sum_{k=0}^j Z_k[i] \right] \\ &= hX[i] + \frac{1}{j+1} \sum_{k=0}^j Z_k[i] \end{aligned}$$

where $Z_k[i]$ are i.i.d gaussian R.V. so they will be added and their variance will be reduced hence SNR will go high. Now we will have new error function where SNR

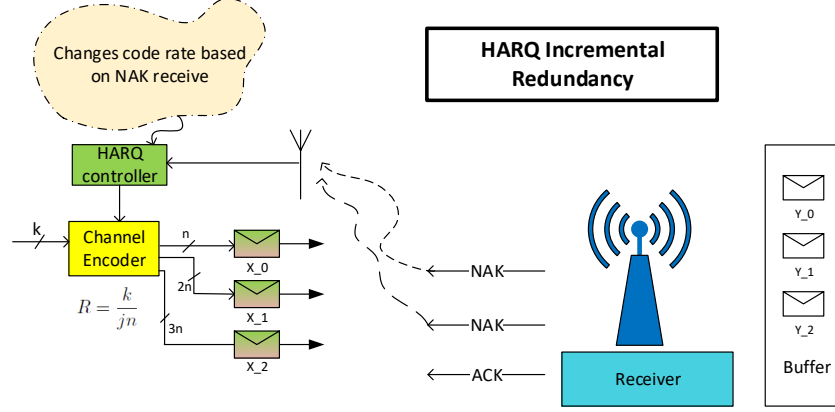


Figure 2-3: Hybrid ARQ Incremental Redundancy Block Diagram

will be the product of number of trials used to decode the message.

$$\varepsilon = Q(k, n, j \times SNR) \quad (2.11)$$

we will use the above (2.11) and (4.14) to minimize (2.7) by maximizing the SNR.

2.4.2 Incremental Redundancy

In Incremental redundancy a packet is re-transmitted with varying parity bits each time a NAK is received by the transmitter. At the receiver side instead of using MRC the receiver tends to decode the current packet and the saved packet on the basis of increased block-length. The new block-length which is jn where j is the number of transmissions results in the decreased code rate. A system model incorporating Incremental redundancy is shown in Figure 2-3. When this scheme is used the re-transmitted message will now be changed with additional encoding and at the receiver we will combine these two different coded messages to form a low code rate packet which has strong error correction capability. The new code rate will be:

$$R = \frac{k}{jn} \quad (2.12)$$

where j is the number of trials use to decode the message. The new error function can be formulated as:

$$\varepsilon = Q(k, jn, SNR) \tag{2.13}$$

we will use (2.13) and (4.14) to minimize 2.7 by utilizing the increased block-length in chapter 4

Chapter 3

Literature Review

In the past decade many people started working on reducing latency for factory automation and low latency communication for wireless networks. There has been two directions involved in order to achieve this. One way is to modify existing control algorithms such that they cope with existing wireless standards and latency introduced by current communication systems and the other is to modify current protocols of communication such as LTE, IEEE 802.11n, Zigbee etc so they can cope with current standards in wired communication (e.g. SERCOS-III, Profibus etc) by providing low latencies and high reliability. The later focuses by complete redesign PHY and MAC layer so that they can meet the tight latency and reliability bounds in a slow fading environment. This chapter explains some previous work that has been done in the literature where people have attempted to solve this problem from modifying existing protocols, using co-operative strategies all the way up to complete design of new protocols.

3.1 Co-operative Strategies

Low latency and high reliability was always a challenging task for a wireless communication engineer. One way to solve this problem is by fixing initial transmission

schedule, budgeting enough time for worst-case number of transmissions and by using very low code rates to optimally balance the number of re-transmissions with coding overhead [13]. [13] have attempted initial approaches in designing a protocol in which a controller sends packets to all the nodes and the nodes failing to decode the packet sends NAK back to controller, after which controller calculates worst-case number of re-transmissions and by changing code rate it re-transmits the packet with an increased probability of decoding. Author uses TDMA to fix the time schedules of a controller to nodes and from nodes to controller re-transmissions. Another technique was proposed in 2015 by the same group was to use co-operative communication for achieving high reliability and low latency communication. They have developed a new communication protocol named "OccupyCoW" (Optimizing Cooperative Communication for Ultra-reliable Protocols Yoking Control Onto Wireless) [14]. OccupyCoW works by increasing diversity and using multiple nodes in co-operation. These nodes simultaneously retransmits to the nodes who have failed the decoding in the first phase of transmission and by using semi-fixed resources and low-rate coding they guarantee that the latency is not increased for test case of industrial printer having 30 nodes in co-operation. Occupy-CoW increases diversity and co-operation among nodes as well as increases complexity and power consumption of each node by making it as a relay which ultimately degrades performance metrics set by M-MTC. In 2016 they have improved performance of Occupy-CoW by using network coding and naming the new protocol as XOR-CoW [15]. Their results indicate that they have achieved round trip time under 2 ms with average throughput of 4.8 Mbps with probability of error up to 10^{-9} for the same case of industrial printer. Where Occupy-CoW uses co-operative strategies on a small environment to achieve low latencies, many people have worked for long distance latency minimization over microwave links and fibre links as well. According to [16] optimal selection of decode-and-forward DF and amplify-and-forward AF relays can reduce the end-to-end transmission delays for long

distance by using cooperative strategies but this approach is not performed for slow fading analysis and have limitations in scenarios of Machine Type Communication.

3.2 Air Interface designs

Reference [17] suggest a good approach to overcome this latency problem by using short transmission intervals without the need of re-transmissions. They have re-designed the system for factory automation case including multiple access techniques, radio access network, backhaul, storage etc. and by using OFDM and diversity techniques with transmission time of $100\mu s$ they achieve latencies up to 1 ms for a packet size of 100 bits. They have limited their transmission for smaller packet sizes which ultimately sets a bound on coverage area as well. Another group which is working to reduce latencies for factory automation scenario suggests using OFDM based 5G radio interface [18]. Their design of new air interface claims to achieve high reliability and sub-millisecond latencies by modifying existing design parameters of LTE, 3GPP based upon OFDM. They suggest using convolution codes instead of turbo codes as they are faster to decode, and further suggests to increase diversity levels and reduced transmission time intervals by using shorter OFDM symbol duration. Their claim depends upon the high availability of coverage in a factory hall deployment which can be achieved by changing parameters such as required SNR, signal bandwidth and number of antenna's required.

Chapter 4

Coding Rate and Latency

In this chapter we will use (2.7) to formulate a series optimization problems which minimizes the Latency function under three cases (ARQ, HARQ-CC, HARQ-INR). We begin by first finding the expression for average no. of packets under ARQ, formulating an optimization problem and finding the bound on R where our optimal solution lies. Later we introduce expression for average no. of channel use for Hybrid ARQ model, formulating its optimization problems for the case of Chase Combining and Incremental Redundancy.

4.1 Latency Formulation under ARQ

Latency function defined in (2.7) requires average number of re-transmissions \mathcal{T} , where for ARQ we find average number of re-transmissions as:

4.1.1 Average no. of transmissions under ARQ

Average number of re-transmissions for a point to point link in ARQ can be found by using probabilistic theory under the assumption of fixed SNR. Suppose the probability of failure to decode message is given by ε which was introduced by [8] in (2.5). Then

the probability of successful transmission is given as $1 - \varepsilon$. Then for a given SNR we can find the average number of re-transmissions by calculating probability of N re-transmissions and then taking expectation to average out the number of transmissions = no of re-transmissions + 1.

$$\begin{aligned} P_N(n \mid \text{SNR}) \\ &= P\{N = n \mid \text{SNR}\} \\ &= \varepsilon^n (1 - \varepsilon) \end{aligned}$$

where ε^n represents the probability of failed decoding for n re-transmissions and $(1 - \varepsilon)$ is the probability of successful re-transmission. Now we limit our number of re-transmission to N_{max} . The reason to limit these re-transmissions is governed by network reliability which sets our constraint as $\varepsilon^{N_{max}} < p^*$. We will denote the new variable as N_p to limit our re-transmissions which is defined below:

$$N_p = \min(N, N_{max}) \tag{4.1}$$

As our primary target is to minimize number of re-transmissions so we will take minimum of N or N_{max} . Now we will find the Probability of re-transmissions w.r.t new variable N_p

$$\begin{aligned} P_{N_p} &= P(N_p = n_p) \\ &= P(\min(N, N_{max}) = n_p) \end{aligned}$$

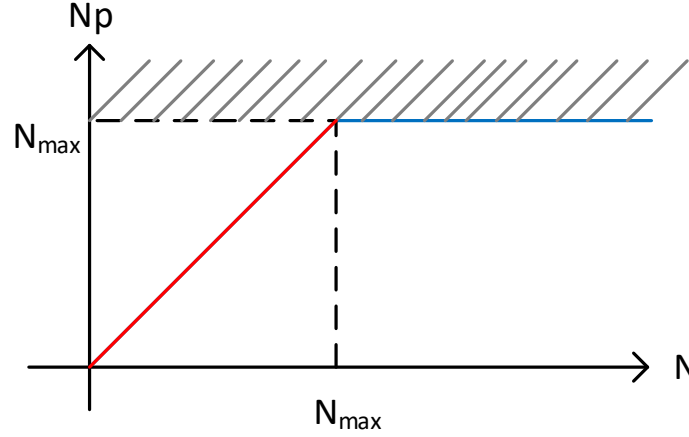


Figure 4-1: Visual Representation of the cases of N_p

This can be further divided into following cases:

$$\begin{cases} P(N = n_p), & n_p < N_{max} \\ 0, & n_p > N_{max} \\ \sum_{i=N_{max}}^{\infty} P_N(i), & n_p = N_{max} \end{cases} \quad (4.2)$$

A visual representation of (4.2) is shown in Figure 4-1 which represents the cases of N_p . The rising curve shows the case where $P(N = n_p)$ when $n_p < N_{max}$. Second case is shown by dashes when $n_p > N_{max}$ then $P_{N_p} = 0$ as there is no region common. The third case is shown by a straight blue line such that when $n_p = N_{max}$ then

$$P_{N_p} = \sum_{i=N_{max}}^{\infty} P_N(i)$$

The first case is same as $P_{N_p} = \varepsilon^{n_p}(1 - \varepsilon)$. The third case is simplified below:

$$\begin{aligned}
P_{N_p} &= \sum_{i=N_{max}}^{\infty} P_N(i) \\
&= 1 - \sum_{i=0}^{N_{max}-1} P_N(i) \\
&= 1 - \sum_{i=0}^{N_{max}-1} \varepsilon^i (1 - \varepsilon) \\
&= 1 - [1 - \varepsilon + \varepsilon - \varepsilon^2 + \varepsilon^2 \dots - \varepsilon^{N_{max}-1} + \varepsilon^{N_{max}-1} - \varepsilon^{N_{max}}] \\
&= 1 - 1 + \varepsilon^{N_{max}} \\
&= \varepsilon^{N_{max}}
\end{aligned}$$

Now we can write (4.2) as:

$$P(N = n_p \mid \text{SNR}) = \begin{cases} \varepsilon^{n_p} (1 - \varepsilon), & n_p < N_{max} \\ 0, & n_p > N_{max} \\ \varepsilon^{N_{max}}, & n_p = N_{max} \end{cases} \quad (4.3)$$

To find the average no of re-transmissions we will take expectation of (4.3)

$$\mathcal{T} = E[N_p \mid \text{SNR}] = \sum_{n=0}^{N_{max}-1} n \varepsilon^n (1 - \varepsilon) + N_{max} \varepsilon^{N_{max}} \quad (4.4)$$

In (4.4) we will now solve the first part i.e. $\sum_{n=0}^{N_{max}-1} n \varepsilon^n (1 - \varepsilon)$

we know the solution of

$$\sum_{n=0}^{N_{max}-1} \varepsilon^n = \frac{1 - \varepsilon^{N_{max}}}{1 - \varepsilon} \quad (4.5)$$

which is a geometric progression. We will take derivative of (4.5) w.r.t ε

$$\begin{aligned}
\frac{d}{d\varepsilon} \sum_{n=0}^{N_{max}-1} \varepsilon^n &= \frac{d}{d\varepsilon} \frac{1 - \varepsilon^{N_{max}}}{1 - \varepsilon} \\
\sum_{n=0}^{N_{max}-1} n \varepsilon^{n-1} &= \frac{1 - \varepsilon^{N_{max}}}{(1 - \varepsilon)^2} + \frac{-N_{max} \varepsilon^{N_{max}-1}}{1 - \varepsilon} \\
&= \frac{(1 - \varepsilon^{N_{max}}) - (1 - \varepsilon)(N_{max} \varepsilon^{N_{max}-1})}{(1 - \varepsilon)^2}
\end{aligned}$$

since the summation in the L.H.S is $\sum_{n=0}^{N_{max}-1} n \varepsilon^{n-1}$ but we require $\sum_{n=0}^{N_{max}-1} n \varepsilon^n$ So we multiply with ε on both sides

$$\begin{aligned}
\sum_{n=0}^{N_{max}-1} n \varepsilon^n &= \frac{\varepsilon[(1 - \varepsilon^{N_{max}}) - (1 - \varepsilon)(N_{max} \varepsilon^{N_{max}-1})]}{(1 - \varepsilon)^2} \\
&= \frac{\varepsilon^{N_{max}+1}(N_{max} - 1) - N_{max} \varepsilon^{N_{max}} + \varepsilon}{(1 - \varepsilon)^2} \tag{4.6}
\end{aligned}$$

we will now use the result of (4.6) in (4.4) :

$$\begin{aligned}
E[N_p \mid \text{SNR}] &= \frac{(1 - \varepsilon)(\varepsilon^{N_{max}+1}(N_{max} - 1) - N_{max} \varepsilon^{N_{max}} + \varepsilon)}{(1 - \varepsilon)^2} + N_{max} \varepsilon \\
&= \frac{(\varepsilon^{N_{max}+1}(N_{max} - 1) - N_{max} \varepsilon^{N_{max}} + \varepsilon)}{1 - \varepsilon} + N_{max} \varepsilon \\
&= \frac{N_{max} \varepsilon^{N_{max}+1} - \varepsilon^{N_{max}+1} - N_{max} \varepsilon^{N_{max}} + \varepsilon + N_{max} \varepsilon^{N_{max}} - N_{max} \varepsilon^{N_{max}+1}}{1 - \varepsilon} \\
&= \frac{\varepsilon - \varepsilon^{N_{max}+1}}{1 - \varepsilon}
\end{aligned}$$

Since we have calculated our average no of re-transmissions the total no of transmis-

sions will be the average no of retransmissions + 1

$$\begin{aligned}\mathcal{T} = E[N_p \mid \text{SNR}] &= \frac{\varepsilon - \varepsilon^{N_{max}+1}}{1 - \varepsilon} + 1 \\ &= \frac{1 - \varepsilon^{N_{max}+1}}{1 - \varepsilon}\end{aligned}\tag{4.7}$$

Equation (4.7) will give us the average number of transmissions in a point to point communication for a fixed SNR.

4.1.2 Optimization problem for ARQ

We would now formulate the optimization problem by using (2.7) where our optimization variables are N_{max} and code rate R .

$$\begin{aligned}\underset{R \geq 0, N_{max}}{\text{minimize}} \quad & E[N_p \mid \text{SNR}] \times \frac{n}{k} \\ \text{subject to} \quad & \varepsilon^{N_{max}+1} \leq p^*\end{aligned}\tag{4.8}$$

Using (4.7) in (4.8) we get:

$$\begin{aligned}\underset{R \geq 0, N_{max}}{\text{minimize}} \quad & \frac{1 - \varepsilon^{N_{max}+1}}{1 - \varepsilon} \times \frac{n}{k} \\ \text{subject to} \quad & \varepsilon^{N_{max}+1} \leq p^*\end{aligned}\tag{4.9}$$

where p^* determines reliability of the system which we want to achieve. Now we modify our constraint as:

$$\begin{aligned}\varepsilon^{N_{max}+1} &\leq p^* \\ \frac{\log(p^*)}{\log(\varepsilon)} &\leq N_{max} + 1 \\ N_{max} &\geq \frac{\log(p^*)}{\log(\varepsilon)} - 1\end{aligned}$$

If N_{max} increases then $\varepsilon^{N_{max}+1}$ decreases which means $(1 - \varepsilon^{N_{max}+1})$ increases. So, to minimize (4.9) we choose $N_{max} = \frac{\log(p^*)}{\log(\varepsilon)} - 1$. This will change the constraint in (4.9) to $\varepsilon^{N_{max}+1} = p^*$. So, the new optimization problem becomes:

$$\underset{R \geq 0}{\text{minimize}} \quad \frac{1 - p^*}{1 - \varepsilon} \times \frac{1}{R} \quad (4.10)$$

4.1.3 Bounds on R for minimizing Latency in ARQ

In section 4.1.2 we have introduced the optimization problem (4.10) for ARQ under the assumption of fixed SNR. In (4.10) p^* is the reliability performance constraint which is known during the transmission. Hence, minimizing (4.10) is same as maximizing $R(1 - \varepsilon)$. So, we have:

$$\underset{R \geq 0}{\text{maximize}} \quad R(1 - \varepsilon) \quad (4.11)$$

To find the solution to (4.11) we first find if the function $R(1 - \varepsilon)$ is convex or concave. In order to find if a function is convex its we compute its second derivative which will turn out to be positive otherwise the function is concave.

$$\frac{d^2}{dR^2} R \left(1 - Q \left[(C - R) \left(\sqrt{\frac{k}{RV}} \right) \right] \right)$$

Since $\frac{k}{R} = n$ and let $\sqrt{\frac{n}{V}} = b$ we have:

$$\begin{aligned}
\frac{d^2}{dR^2} R(1 - \varepsilon) &= \\
&= \frac{d^2}{dR^2} R(1 - Q[(C - R)b]) \\
&= \frac{d^2}{dR^2} R \left[Q \left(\frac{R - C}{\frac{1}{b}} \right) \right] \\
&= \frac{d^2}{dR^2} R [1 - F_{R'}(R)] \\
&= \frac{d^2}{dR^2} [R - RF_{R'}(R)] \\
&= \frac{d^2}{dR^2} R - \frac{d^2}{dR^2} RF_{R'}(R) \\
&= 0 - \frac{d}{dR} [F_{R'}(R) + Rf_{R'}(R)] \\
&= -\frac{d}{dR} F_{R'}(R) - \frac{d}{dR} Rf_{R'}(R) \\
&= -f_{R'}(R) - f_{R'}(R) - R\frac{d}{dR} f_{R'}(R)
\end{aligned}$$

$$\begin{aligned}
\frac{d^2}{dR^2} R(1 - \varepsilon) &= -2f_{R'}(R) - R\frac{d}{dR} \left[\frac{b}{\sqrt{2\pi}} \exp \left(\frac{-(R - C)^2}{\frac{2}{b^2}} \right) \right] \\
&= -2f_{R'}(R) - R \left[\frac{b}{\sqrt{2\pi}} \exp \left(\frac{-(R - C)^2}{\frac{2}{b^2}} \right) \left(\frac{-2(R - C)}{\frac{2}{b^2}} \right) \right] \\
&= -2f_{R'}(R) - R \left[f_{R'}(R) \left(\frac{-(R - C)}{\frac{1}{b^2}} \right) \right] \\
&= -2f_{R'}(R) - Rf_{R'}(R) \left(\frac{-(R - C)}{\frac{1}{b^2}} \right) \\
&= f_{R'}(R) \left[-2 - R \left(\frac{-(R - C)}{\frac{1}{b^2}} \right) \right] \\
&= f_{R'}(R) [-2 + R^2b^2 - RCb^2]
\end{aligned} \tag{4.12}$$

where (4.12) is the second derivative of optimization problem (4.11) For second derivative to be positive i.e. convex or negative i.e. concave, we can see that its first part

$f_{R'}(R)$ is always positive and second part of the expression $[-2 + R^2b^2 - RCb^2]$ is a quadratic function which is convex. To analyze this function we first find its zero crossing and find the region in which this function is negative. We solve the second term for R

$$R^2b^2 - RCb^2 - 2 = 0$$

$$R = \frac{Cb^2 \pm \sqrt{C^2b^4 + 8b^2}}{2b^2}$$

Now we know that for two values of R the second derivative becomes 0 i.e. $R = \frac{Cb^2 + \sqrt{C^2b^4 + 8b^2}}{2b^2}$ and $R = \frac{Cb^2 - \sqrt{C^2b^4 + 8b^2}}{2b^2}$ and since the second term is a convex function then for $\frac{Cb^2 - \sqrt{C^2b^4 + 8b^2}}{2b^2} < R < \frac{Cb^2 + \sqrt{C^2b^4 + 8b^2}}{2b^2}$ it will be negative and hence the second derivative will also be negative which means the (4.11) will become concave function. where we can also see that:

$$R = \frac{Cb^2 \pm \sqrt{C^2b^4 + 8b^2}}{2b^2}$$

$$R = \frac{Cb^2 \left[1 \pm \sqrt{1 + \frac{8b^2}{C^2b^4}} \right]}{2b^2}$$

$$R = \frac{Cb^2 \left[1 \pm \sqrt{1 + \frac{8}{C^2b^2}} \right]}{2b^2}$$

Since $\frac{8}{C^2b^2}$ is always positive quantity so $Cb^2 < \sqrt{C^2b^4 + 8b^2}$. Hence we can say that $R \neq \frac{Cb^2 - \sqrt{C^2b^4 + 8b^2}}{2b^2}$ as it will make $R < 0$ which is not possible. On the other hand we can see that the second derivative of optimization problem (4.11) is negative only when $\frac{Cb^2 - \sqrt{C^2b^4 + 8b^2}}{2b^2} \leq R \leq \frac{Cb^2 + \sqrt{C^2b^4 + 8b^2}}{2b^2}$ but since R can't take the negative value so the new range becomes:

$$0 \leq R \leq \frac{Cb^2 + \sqrt{C^2b^4 + 8b^2}}{2b^2} \quad (4.13)$$

This is the range in which the second derivative is always negative and hence the function $R(1 - \varepsilon)$ becomes concave and only one maxima will exist. Here we make an argument that the function described in (4.11) is a decreasing function as $R \rightarrow \infty$ because $(1 - \varepsilon)$ decreases exponentially as compared to linearly increasing R . Hence, in solving (4.10) we will look for R in the range specified in (4.13) in which only one solution will exist as the function is concave in that region. Numerical methods like bi-section method can be applied in solving (4.13). Table 5.1 verifies our result in (4.13) and simulation results of (4.11) are shown in Figure 5-1.

4.2 Latency Formulation under Hybrid ARQ

In Chapter 2 we briefly explained the mechanism of Hybrid ARQ. In this section we will explain mathematical formulation for average number of re-transmissions under Hybrid ARQ and formulation of optimization problem both for Chase Combining and Incremental Redundancy.

4.2.1 Average no. of channel uses in Hybrid ARQ

According to [19] the average number of channel uses to decode a packet under HARQ is given by:

$$\mathcal{T} = \sum_{m=1}^{N_{max}} n_m \varepsilon_{m-1} \quad (4.14)$$

where we define ε_m as the probability that the message is not decoded up to the end of the m -th transmission and $\varepsilon_0 = 0$ and n_m is the length of sub-codeword where

$$n_{N_{max}} = \sum_{m=1}^{N_{max}} n_m$$

Lemma 4.2.1. *Average no. of channel uses in Hybrid ARQ*

Proof. Let E_i be the event of failed decoding. Then: $\varepsilon_m = Pr(E_1 \cap E_2 \cap \dots E_{m-1} \cap E_m)$ and $\varepsilon_{i-1} - \varepsilon_i$ be the probability that the packet will be decoded in exactly i transmissions.

$$\begin{aligned}
\mathcal{T} &= \sum_{i=1}^{N_{max}-1} \left(\sum_{j=1}^i n_j \right) Pr(E_1 \cap E_2 \dots E_i^c) + \left(\sum_{j=1}^{N_{max}} n_j \right) Pr(E_1 \cap E_2 \dots E_M) \\
&= \sum_{i=1}^{N_{max}-1} \sum_{j=1}^i n_j (\varepsilon_{i-1} - \varepsilon_i) + \left(\sum_{i=1}^{N_{max}} n_j \right) \varepsilon_{N_{max}-1} \\
&= \sum_{i=1}^{N_{max}-1} \sum_{j=1}^i n_j \varepsilon_{i-1} - \sum_{i=1}^{N_{max}-1} \sum_{j=1}^i n_j \varepsilon_i + \left(\sum_{i=1}^{N_{max}} n_j \right) \varepsilon_{N_{max}-1} \\
&= \sum_{i=1}^{N_{max}-1} \sum_{j=1}^i n_j \varepsilon_{i-1} - \sum_{i=2}^{N_{max}} \sum_{j=1}^{i-1} n_j \varepsilon_{i-1} + \sum_{j=1}^{N_{max}} n_j \varepsilon_{N_{max}-1} \\
&= \sum_{j=1}^1 n_j \varepsilon_0 + \sum_{i=2}^{N_{max}-1} \left[\sum_{j=1}^i n_j \varepsilon_{i-1} - \sum_{j=1}^{i-1} n_j \varepsilon_{i-1} \right] - \sum_{j=1}^{N_{max}-1} n_j \varepsilon_{N_{max}-1} + \sum_{j=1}^{N_{max}} n_j \varepsilon_{N_{max}-1} \\
&= n_1 \varepsilon_0 + \sum_{i=2}^{N_{max}-1} n_i \varepsilon_{i-1} + n_{N_{max}} \varepsilon_{N_{max}-1} \\
&= \sum_{i=1}^{N_{max}} n_i \varepsilon_{i-1}
\end{aligned}$$

□

4.2.2 Optimization problem for Chase Combining

In chase combining since we utilize MRC which increases SNR of the received packet as explained in 2.4.1 which ultimately results in the increased capacity by:

$$C_m = \log_2(1 + mP) \quad (4.15)$$

where m is the number of re-transmissions. We will utilize (2.11) in (4.14) to formulate optimization problem to minimize (2.7) where in chase combining block-length will

remain fixed with each re-transmission. The optimization problem can be formulated as:

$$\begin{aligned}
& \underset{R \geq 0, N_{max}}{\text{minimize}} && \sum_{m=1}^{N_{max}} \varepsilon_{m-1} \times \frac{n}{k} \\
& \text{subject to} && \varepsilon_{N_{max}} \leq p^*
\end{aligned} \tag{4.16}$$

Using results of [8] we modify ε for chase combining as:

$$\varepsilon = Q \left[(C_m - R) \sqrt{\left(\frac{n}{V_m} \right)} \right] \tag{4.17}$$

Though in (4.17) we have used V_m instead of V but for high SNR's V_m approaches 1 and hence it has negligible effect. Using (4.17) in (4.16) the final optimization problem for chase combining comes out to be:

$$\begin{aligned}
& \underset{R \geq 0, N_{max}}{\text{minimize}} && \sum_{m=1}^{N_{max}} Q \left[(C_{m-1} - R) \sqrt{\left(\frac{n}{V_{m-1}} \right)} \right] \times \frac{1}{R} \\
& \text{subject to} && Q \left[(C_{N_{max}} - R) \sqrt{\left(\frac{n}{V_{N_{max}}} \right)} \right] \leq p^*
\end{aligned} \tag{4.18}$$

Closed form solution and bounds on R for (4.18) are harder to analyze so we present simulation results in chapter 5.

4.2.3 Solution to N_{max} under Chase Combining

In (4.16) we have the constraint $\varepsilon_{N_{max}} \leq p^*$ where we can find semi-analytical lower bound on N_{max} by ignoring the effects of channel dispersion V given as:

$$\begin{aligned}
Q \left[(C_{N_{max}} - R) \sqrt{n} \right] &\leq p^* \\
\left[(\log_2(1 + N_{max}P) - R) \sqrt{n} \right] &\geq Q^{-1}(p^*) \\
\log_2(1 + N_{max}P) &\geq \frac{Q^{-1}(p^*)}{\sqrt{n}} + R \\
N_{max} &\geq \frac{2^{\left(\frac{Q^{-1}(p^*)}{\sqrt{n}} + R \right)} - 1}{P}
\end{aligned} \tag{4.19}$$

In (4.18) our objective function is an increasing function of N_{max} and minimizing (4.18) requires N_{max} to be as smallest as possible. Hence, we will have the following expression of N_{max} for Chase Combining

$$N_{max} \simeq \left\lceil \frac{2^{\left(\frac{Q^{-1}(p^*)}{\sqrt{n}} + R \right)} - 1}{P} \right\rceil \tag{4.20}$$

Where ceil is performed as $N_{max} \in \mathbf{Z}^+$.

4.2.4 Optimization problem for Incremental Redundancy

As explained in 2.4.2 the block-length increases with each re-transmission. Hence, using results of [8] and [19] the error function for HARQ-INR is given by:

$$\varepsilon = Q \left[\left(C - \frac{R}{m} \right) \sqrt{\left(\frac{mn}{V} \right)} \right] \tag{4.21}$$

where in our model the transmitter sends the same message length n with changed parity bits every time, the receiver combines the received message and the buffered message such that the entire packet length is increased and hence the effective code

rate is reduced. Similar to Chase Combining the optimization problem for HARQ-INR is formed as:

$$\begin{aligned}
& \underset{R \geq 0, N_{max}}{\text{minimize}} && \sum_{m=1}^{N_{max}} Q \left[\left(C - \frac{R}{m-1} \right) \sqrt{\left(\frac{(m-1)n}{V} \right)} \right] \times \frac{1}{R} \\
& \text{subject to} && Q \left[\left(C - \frac{R}{N_{max}} \right) \sqrt{\left(\frac{(N_{max}) \times n}{V} \right)} \right] \leq p^*
\end{aligned} \tag{4.22}$$

Closed form solution and bounds on R for (4.22) are harder to analyze, though our conjecture is that the solution to (4.22) lies at $R \rightarrow \infty$. Simulated results of (4.22) shown in Chapter 5 verifies with our conjecture.

Chapter 5

Simulation Results

This chapter introduces simulation results to the problems formulated in Chapter 4.

5.0.1 Performance of ARQ at different SNR

In ARQ, we have calculated bound on code rate in section 4.1.3 where the result is given in (4.13). Figure 5-1 shows simulated result of (4.11). It is clear from Figure 5-1 that for high SNR, $R(1 - \varepsilon)$ is maximum; Hence we will have a minimum latency, and for low SNR we will have high a latency. Table 5.1 verifies that the optimal code rate for results shown in Figure 5-1 lies within the bound computed in (4.13).

Table 5.1: Optimal Code Rate for different SNR in ARQ

SNR (dB)	Optimal Latency	Optimal R	Upper Bound on R
-2	2.64	0.56	0.86
-1	2.11	0.66	0.99
0	1.70	0.76	1.14
1	1.37	0.91	1.30
2	1.13	1.06	1.49

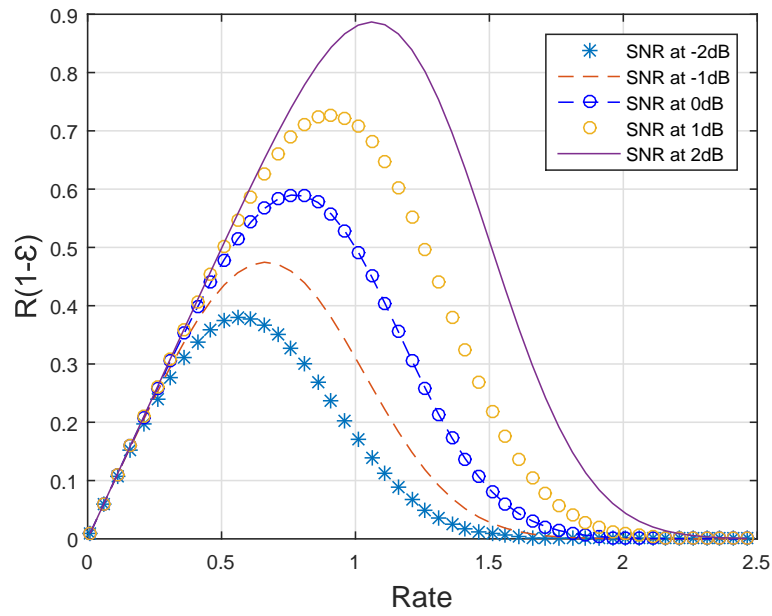


Figure 5-1: Performance of ARQ at different SNR

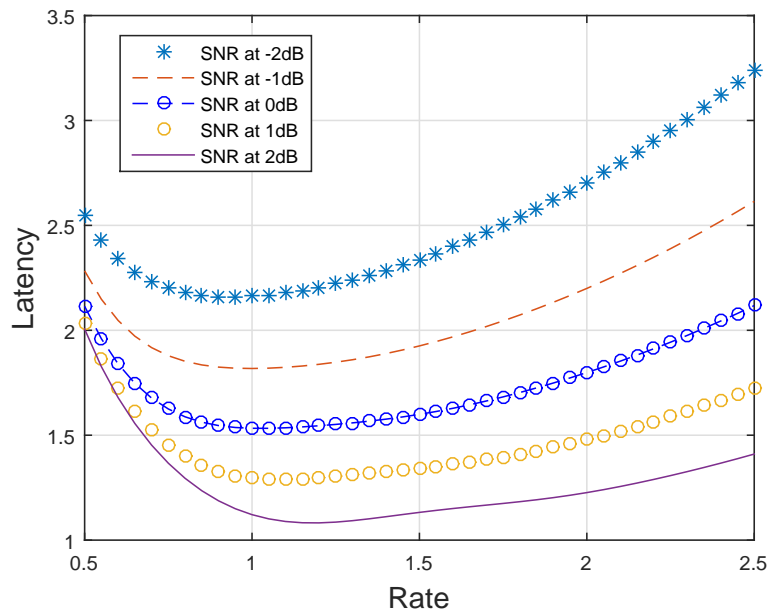


Figure 5-2: Performance of Chase Combining at different SNR

Table 5.2: Optimal Latency & Code Rate for different SNR in Chase Combining

SNR (dB)	Optimal Latency	Optimal R
-2	2.16	0.95
-1	1.82	1
0	1.53	1.05
1	1.29	1.10
2	1.08	1.20

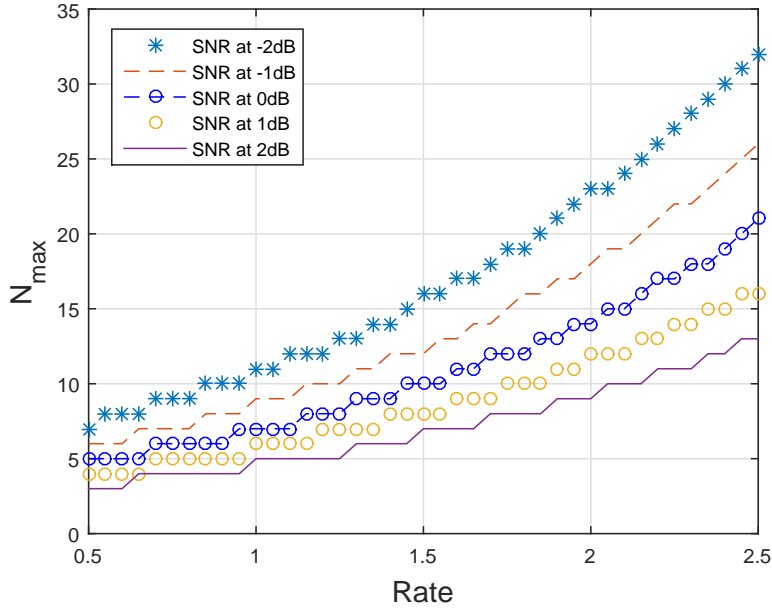


Figure 5-3: N_{max} in Chase Combining at different SNR

5.0.2 SNR vs Latency comparison under Chase Combining

In Figure 5-2, we can see that whenever SNR is high chase combining performs better and achieve lower latencies. Table 5.2 shows optimal values of latencies after computing optimal code rate at different SNR. Comparing Table 5.1 and Table 5.2 we can see that for the same block-length Chase Combine performs better than ARQ in a lower SNR regime, providing low latencies at higher code rates thus reducing complexity.

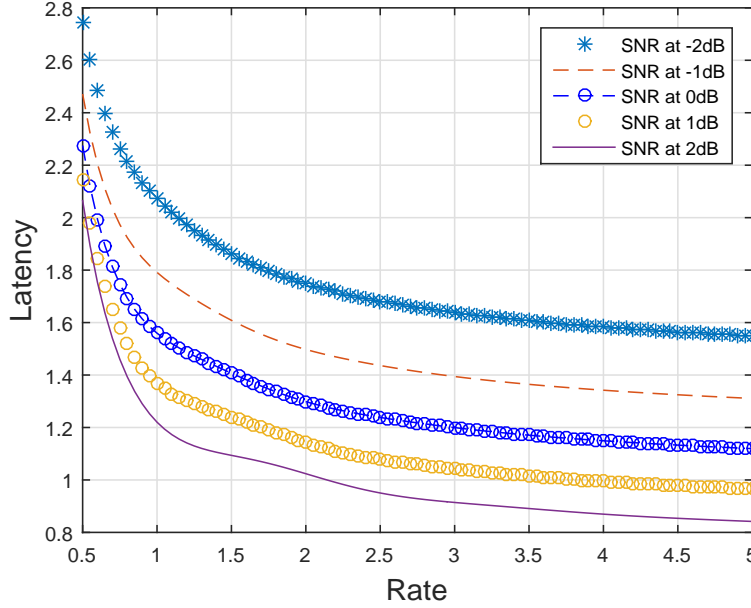


Figure 5-4: Performance of Incremental Redundancy at different SNR

In Figure 5-3 lower SNR requires more number of re-transmissions at higher code rate, which causes degradation in performance. Hence, there is a trade-off between achieving lower latencies with lower SNR using Chase Combining as we require more and more no. of re-transmissions. Since, N_{max} will always be a integer, we can observe step rises in Figure 5-3. These steps will cause transitions in Figure 5-2 at higher blocklengths.

5.0.3 Performance of Incremental Redundancy at different SNR

Figure 5-4 shows that optimal solution to latency will always lie when R approaches ∞ . Simulation result shows that for Incremental Redundancy N_{max} increases linearly with code rate R . Though analytical solution to Incremental case is harder to compute we provide intuitive explanation for this case. So, whenever a packet is re-transmitted the receiver stores the packet in its memory buffer, let the buffer maximum capacity be $s_{max} = 10$ (where we don't always know the size of buffer at the receiver) and

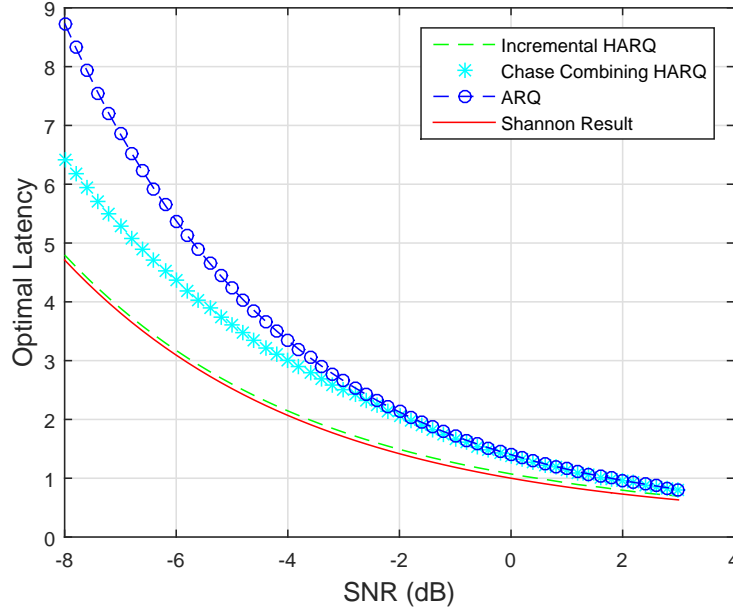


Figure 5-5: Optimal Latency vs SNR for all Schemes

we set the packet length $n = 3$ such that whenever the buffer is full the receiver would be able to decode the packet. If the buffer size reaches $s = 9$ then on next transmission the buffer will be over-flown, and its size would be $s = 12$ where the overflow is an increase in latency. Hence, in order to minimize latency we would decrease the packet length and upon decreasing n we are increasing R hence the solution for optimal latency will always lie at $R \rightarrow \infty$.

5.0.4 Performance analysis of Hybrid ARQ vs ARQ

This section explains the performance comparison in the form of spectral efficiency and optimal latency for different SNR for Hybrid ARQ schemes (Chase Combining & Incremental Redundancy) vs ARQ and Shannon's Result. Figure 5-5 shows that HARQ-INR will achieve lower latencies than all other schemes and performs close to Shannon's theoretical results, the reason for this is the performance metric in Incremental Redundancy which is the block-length n . As n increases with every re-

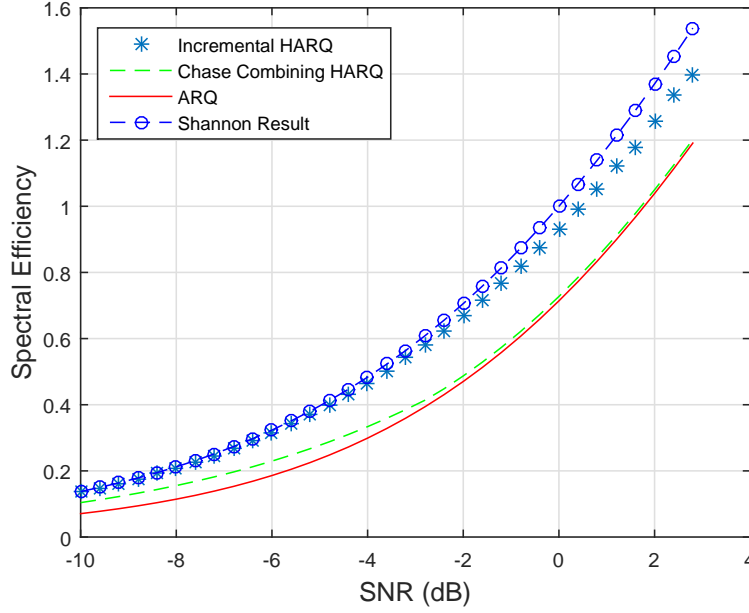


Figure 5-6: Spectral Efficiency vs SNR for all Schemes

transmission this will make it perform better than all other schemes. ARQ and Chase Combining have comparable results at higher SNR's but Chase combining performs superior to ARQ at extreme lower SNR's. Hence, Hybrid ARQ has an advantage over traditional ARQ. In Figure 5-6 we can see the graph of spectral efficiency which is the expected no. of bits per channel use vs SNR. It is clear that at lower SNR's HARQ-INR and Shannon's result are extremely closer to each other whereas at higher SNR's Shannon's theocratical result performs superior to Incremental Redundancy as HARQ-INR is still bounded by a fixed re-transmission of packet length. Whereas in low SNR regime, Chase Combining will have much better spectral efficiency than ARQ though its performance is still inferior as compared to HARQ-INR.

Chapter 6

Conclusion and Future Work

In this thesis, we have studied the effects of code rate on latency under a finite block-length regime for three schemes. Under finite block-length and finite no of channel uses one finds trade-offs associated in achieving high reliability, ultra-low latencies and rates equal to maximum capacity of the channel at the same time. By using expressions and bounds calculated by [4] over maximal rate under finite block-length, we analyze the solution to low latency and high reliability demand by formulating a series of optimization problems by fixing block-length, while constraining under network reliability p^* and setting code rate R and maximal number of transmissions N_{max} as our objective variables such that our average number of packets are reduced for three cases (a). Automatic Repeat Requests (ARQ), (b) Hybrid Automatic Repeat Requests - Chase Combining (HARQ-CC), and (c) Hybrid Automatic Repeat Requests -Incremental Redundancy (HARQ-INR). In Chapter 4 we derived the analytical expressions for average number of channels uses for ARQ and HARQ cases, and formulating optimization problems by constraining under the performance reliability of the network which is usually 10^{-9} as required by tactile internet and 5G. We further computed simulated and semi-analytical results which gives us bounds on code rate R for an ARQ case and in which for a given SNR and network reliability we can find

optimal solution to minimal latency. Our simulated results for Hybrid ARQ cases shows that HARQ-INR will always perform better in all of these three schemes and HARQ-CC will always perform better than ARQ especially in the lower SNR regime. Since HARQ-INR relies on increase in block-length with each re-transmission, our result also verifies that HARQ-INR will reach Shannon's approximations under the limit of infinite no of channel uses.

As a future direction, we can optimize over both the finite block-length and average no of packets together which will give us better results in reducing average no. of channel uses per bit. We have fixed our attention to an assumption that state of the channel is known, whereas in practical scenarios channel needs to be estimated with every transmission. So, as a future scope of our research, minimizing latencies over dynamic channels looks promising and by using different coding schemes we can develop a complete protocol which optimally finds the best R under finite block-length such that the overall round-trip time of a communication system is minimized while achieving ultra-high reliability.

Bibliography

- [1] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, “Fundamental tradeoffs among reliability, latency and throughput in cellular networks,” in *Globecom Workshops (GC Wkshps), 2014*. IEEE, 2014, pp. 1391–1396.
- [2] C. E. Shannon, “A mathematical theory of communication,” *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.
- [3] G. C. Burdea and P. Coiffet, *Virtual reality technology*. John Wiley & Sons, 2003, vol. 1.
- [4] Y. Polyanskiy, *Channel coding: non-asymptotic fundamental limits*. Princeton University, 2010.
- [5] K. Bilstrup, E. Uhlemann, E. G. Strom, and U. Bilstrup, “Evaluation of the ieee 802.11 p mac method for vehicle-to-vehicle communication,” in *Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th*. IEEE, 2008, pp. 1–5.
- [6] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka *et al.*, “Scenarios for 5g mobile and wireless communications: the vision of the metis project,” *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, 2014.

- [7] Y. Polyanskiy, H. V. Poor, and S. Verdú, “Dispersion of gaussian channels,” in *2009 IEEE International Symposium on Information Theory*. IEEE, 2009, pp. 2204–2208.
- [8] —, “Channel coding rate in the finite blocklength regime,” *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.
- [9] P. Frenger, S. Parkvall, and E. Dahlman, “Performance comparison of harq with chase combining and incremental redundancy for hsdpa,” in *Vehicular Technology Conference, 2001. VTC 2001 Fall. IEEE VTS 54th*, vol. 3. IEEE, 2001, pp. 1829–1833.
- [10] A. Chelli and M.-S. Alouini, “Performance of hybrid-arq with incremental redundancy over relay channels,” in *Globecom Workshops (GC Wkshps), 2012 IEEE*. IEEE, 2012, pp. 116–121.
- [11] L. Szczecinski, C. Correa, and L. Ahumada, “Variable-rate transmission for incremental redundancy hybrid arq,” in *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*. IEEE, 2010, pp. 1–5.
- [12] C. Shen, T. Liu, and M. P. Fitz, “On the average rate performance of hybrid-arq in quasi-static fading channels,” *IEEE Transactions on Communications*, vol. 57, no. 11, 2009.
- [13] M. Weiner, M. Jorgovanovic, A. Sahai, and B. Nikolić, “Design of a low-latency, high-reliability wireless communication system for control applications,” in *2014 IEEE International Conference on Communications (ICC)*. IEEE, 2014, pp. 3829–3835.
- [14] V. N. Swamy, S. Suri, P. Rigge, M. Weiner, G. Ranade, A. Sahai, and B. Nikolić, “Cooperative communication for high-reliability low-latency wireless control,” in

- 2015 IEEE International Conference on Communications (ICC)*. IEEE, 2015, pp. 4380–4386.
- [15] V. N. Swamy, P. Rigge, G. Ranade, A. Sahai, and B. Nikolić, “Network coding for high-reliability low-latency wireless control,” in *Wireless Communications and Networking Conference (WCNC), 2016 IEEE*. IEEE, 2016, pp. 1–7.
- [16] I. Marić, “Low latency communications,” in *Information Theory and Applications Workshop (ITA), 2013*. IEEE, 2013, pp. 1–6.
- [17] N. A. Johansson, Y.-P. E. Wang, E. Eriksson, and M. Hessler, “Radio access for ultra-reliable and low-latency 5g communications,” in *2015 IEEE International Conference on Communication Workshop (ICCW)*. IEEE, 2015, pp. 1184–1189.
- [18] O. N. Yilmaz, Y.-P. E. Wang, N. A. Johansson, N. Brahmi, S. A. Ashraf, and J. Sachs, “Analysis of ultra-reliable and low-latency 5g communication for a factory automation use case,” in *Communication Workshop (ICCW), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1190–1195.
- [19] B. Makki, T. Svensson, and M. Zorzi, “Finite block-length analysis of the incremental redundancy harq,” *IEEE Wireless Communications Letters*, vol. 3, no. 5, pp. 529–532, 2014.