Bits through Time

Thèse N° 9319

Présentée le 8 mars 2019

à la Faculté informatique et communications
Laboratoire de théorie de l'information
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Elie NAJM

Acceptée sur proposition du jury

Dr N. Macris, président du jury
Prof. E. Telatar, directeur de thèse
Prof. R. Yates, rapporteur
Prof. A. Ephremides, rapporteur
Prof. P. Thiran, rapporteur

2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Thesis No. 9319 (February 2019)

Thesis presented to the faculty of computer and communication sciences for obtaining the degree of Docteur ès Sciences

Accepted by the jury:

Prof. Emre Telatar
Thesis director

Prof. Anthony Ephremides
Expert

Prof. Pactrick Thiran
Expert

Prof. Roy Yates
Expert

Dr. Nicolas Macris
President of the jury

Ecole Polytechnique Fédérale de Lausanne, 2019

Dedicated to the memory of my grandfather Salem Elias,
(1928-2016)

To my parents and my brother...

تَودُّون ان تَقيسوا الزمن الذي لا يُقاس و لا يُحَدّ،
لكن ما هو خالدٌ فيكم يُدرك انّ الحياة لا يَحُدُّها زمان،
و يعلم انّ الأمس ما هو إلاّ ذاكِرَة اليوم، و أنّ الغد ما هو إلاّ حُلمه،
أليس الزمن كالحبّ نفسه، لا ينقسم و لا يُقاس؟
—جبران خليل جبران، النبي

You would measure time the measureless and the immeasurable.
Yet the timeless in you is aware of life's timelessness,
And knows that yesterday is but today's memory and
tomorrow is today's dream.
And is not time even as love is, undivided and paceless?
— Gibran Khalil Gibran, The Prophet

Abstract

In any communication system, there exist two dimensions through which the information at the source becomes distorted before reaching the destination: the noisy channel and time. Messages transmitted through a noisy channel are susceptible to modification in their content, due to the action of the noise of the channel. Claude E. Shannon, in his seminal paper of 1948 "A Mathematical Theory of Communication", introduces the *bit* as a unit of measure of information, and he lays down the theoretical foundations needed to understand the problem of sending bits reliably through a noisy channel. The distortion measure, which he used to quantify reliability, is the error probability. In his paper, Shannon shows that any channel is characterized by a number that he calls *capacity*: It represents the highest transmission rate that can be used to communicate information with, through this same channel, while guaranteeing a negligible error probability. These foundations led to the development of the field of information theory.

Whereas, even if the messages are sent through a perfect channel, the time they take to reach their destination causes the receiver to acquire a distorted view of the status of the source that generated these messages. For instance, take the case of a monitor interested in the status of a distant process. A sender observes this process and, to keep the monitor up-to-date, sends updates to it. However, if, at any time t, the last received update at the monitor was generated at time $u(t)$, then the information at the receiver reflects the status of the process at time $u(t)$, not at time t. Hence, the monitor has a distorted version of reality. In fact, it has an obsolete version with an *age* of $t - u(t)$. This concept is common in astronomy where, due to the extremely long distances between the observed process and the monitor (Earth), the reality that astronomers have about the status of distant stars corresponds to the state of these stars millions or billions years ago. This is a clear illustration of how the information (or bits) that a monitor seeks can be distorted by time. This example shows the importance, from the receiver point of view, of building a communication system that maintains the monitor up-to-date. In other words, the monitor needs a system that ensures a negligible *age*.

The concept of *age* as a distortion measure in communication systems was first used in 2011 by Kaul et al., in order to assess the performance of a given vehicular network. The aim of the authors was to come up with a transmission scheme that would minimize an age-related metric: the *average age*. Since then, a growing body of works has used this metric to evaluate the performance of multiple communication systems; some of them being subject to resource allocation constraints. The drive behind this interest lies in the importance that status-update applications are gaining

in today's life (in vehicular networks, warehouse and environment surveillance, news feed, etc.).

In this thesis, we choose age as a distortion measure and derive expressions for the *average age* and the *average peak-age* (another age-related metric) for different communication systems. Therefore, we divide this dissertation into two parts: In the first part, we assume that the the updates are transmitted through a noiseless channel that has a random service time. In the second part, we consider a special category of noisy channels, namely the erasure channel. In the first part of this thesis, in order to compute the age-related metrics, we employ queue-theoretic concepts. We study and compare the performance of various transmission schemes under different settings: (*i*) when we are interested in only one source and the service time is gamma distributed, (*ii*) when we are interested in multiple sources with equal priority and, (*iii*) when the sources are assigned different priorities. We show that the optimal transmission scheme when the monitor is interested in a single source loses its optimality when another source of higher priority shares the system.

In the second part of this thesis, we introduce, in our age calculations, the distortion caused by the erasure channel on the transmitted updates. In order to combat the erasures of the channel, we first consider two flavors of the hybrid automatic repeat request (HARQ), an error-correcting protocol implemented in cellular systems: the infinite incremental redundancy (IIR) HARQ and the fixed redundancy (FR) HARQ. We compute, under two different transmission schemes, the *average age* for both IIR and FR; and we show that IIR leads always to a lower *average age* compared to FR. Finally, we focus on the optimal *average age* that could be achieved over an erasure channel. We prove that if the channel-input alphabet is identical to that of the source, then no error coding is needed to achieve the optimal *average age* for which we compute the closed-form expression. However, when the two alphabets are different, we use a random-coding argument to tightly bound the optimal *average age*.

Keywords: Average age, average peak-age, renewal processes, Poisson processes, random codes, HARQ, ergodic theory, erasure channel, Markov chains.

Résumé

Dans tout système de communication, deux facteurs distordent l'information à la source avant sa réception par le destinataire : le canal à bruit et le temps. En un premier temps, les messages transmis á travers un canal á bruit sont susceptibles de voir leur contenu modifié suite á l'action du bruit. Claude E. Shannon introduit, dans son article fondateur de 1948 intitulé « Mathematical Theory of Communication », la notion de *bit* comme unité de mesure de l'information et pose les bases théoriques nécessaires á la compréhension du problème de transmission sûre de bits à travers un canal à bruit. Shannon définit la probabilité d'erreur comme étant la mesure de distorsion nécessaire pour évaluer la sûreté d'un système de communication. En fait, Shannon prouve que chaque canal se charactérise par un nombre qu'il nomme *capacité* et qui représente le débit de transmission maximal qu'aucun peut adopter tout en étant sûr de pouvoir garantir une probabilité d'erreur négligeable. Ces fondations causèrent le développement du domaine de la Théorie de l'Information.

En un second temps, même si les messages (ou updates) sont transmis à travers un canal parfait, l'intervalle de temps que ces updates mettent à arriver á leur destination cause le récepteur de constituer une image tordue de l'état de la source qui génère ces messages. Par example, prenons le cas d'un moniteur intéressé par l'état d'un processus éloigné. Un transmetteur observe ce processus et transmet des updates au moniteur pour le garder á jour. Cependant, si à un certain moment t, le dernier update reçu par le destinataire était généré à l'instant $u(t)$ alors l'information que le moniteur détient ne reflète l'état de la source telle que qu'elle était à l'instant $u(t)$. De là, le moniteur possède une version tordue de la réalité. En effet, ce dernier possède une version obsolète *âgée* de $t - u(t)$. Cette notion est récurrente en astronomie où la réalité que les astronomes peuvent reconstruire à propos d'une étoile éloignée correspond à l'état de l'étoile telle qu'elle se portait il y a des millions voire des milliards d'années. Dans ce cas, la distorsion est surtout dûe aux grandes distances séparant le processus observé (les étoiles) du moniteur (la Terre). Ceci présente une illustration claire sur l'effet que le temps exerce sur l'information demandée par un moniteur. Cet example souligne aussi l'importance, du point de vue du récepteur, de construire un système de communication qui garantit une mise à jour efficace du moniteur. En d'autres termes, le moniteur a besoin d'un système qui puisse assurer un *âge* minimal.

La notion d'*âge*, en tant que mesure de distorsion dans les systèmes de communication, fut utilisée en premier par Kaul et al. en 2011 afin d'évaluer la performance d'un certain réseau véhiculaire. Le but des auteurs consistait à trouver un schéma de transmission qui minimize l'*âge moyen*. Depuis, un nombre croissant d'articles

utilisent cette nouvelle mesure pour évaluer la performance de différents systèmes de communication dont certains sont sujets à des contraintes sur les ressources disponibles. L'intérêt accru envers ce sujet vient de l'importance croissante que les applications de mise à jour gagnent dans notre vie d'aujoud'hui (tels les réseaux véhiculaires, la surveillance des dépôts et de l'environnement, les sources de nouvelles, etc.).

Dans cette thèse, nous choisissons l'âge comme mesure de distorsion et dérivons des expressions pour l'*âge moyen* ainsi que l'*âge de pointe moyen* pour différents systèmes de communications. La thèse se divise en deux parties : dans la première partie nous supposons que les updates sont transmis à travers un canal sans bruit mais avec un temps de service aléatoire. Par contre, dans la seconde partie, nous considérons une catégorie spéciale de canaux à bruit, celle des canaux d'effacements. Dans la première partie de cette dissertation, nous employons des notions empruntées à la théorie des queues afin de calculer les deux mesures d'âge mentionnées précédemment. Nous étudions et comparons les performances d'une variété de schémas de communication sous différents angles : (*i*) lorsque le moniteur n'est intéressé que par une seule source et que le temps de service possède une distribution gamma, (*ii*) lorsque le moniteur est intéressé par plusieurs sources avec des priorités similaires et (*iii*) lorsque les sources possèdent des priorités différentes. Nous démontrons que le schéma optimal de transmission dans le cas d'une source unique devient sous-optimal lorsqu'une nouvelle source avec une priorité plus élevée vient partager le système.

Dans la seconde partie de cette dissertation, nous prenons en considération dans nos calculs la distorsion apportée par le canal d'effacements. Afin de combattre les effacements du canal, nous considérons en un premier temps deux versions du protocol de *hybrid automatic repeat request (HARQ)* qui est implémenté dans les systémes cellulaires : la redondance incrémentale infinie (RII) HARQ et la redondance fixe (RF) HARQ. Nous calculons pour deux schémas de transmission différents les *âges moyens* relatifs à la RII et à la RF. Nous démontrons que la RII présente toujours un *âge moyen* plus petit que celui réalisé par la RF. Finalement, nous nous concentrons sur l'*âge moyen* optimal qui puisse être atteint en présence d'un canal d'effacements. Nous prouvons que si l'alphabet de la source est identique à celui de l'entrée du canal, alors le codage de canal s'avère inutile pour atteindre l'*âge moyen* dont nous dérivons la forme exacte. Toutefois, lorsque les deux alphabets sont différents, nous utilisons un argument basé sur le codage aléatoire afin de trouver des bornes serrées sur l'*âge moyen* optimal.

Mots-clés : Âge moyen, âge de pointe moyen, processus de renouvellement, processus de Poisson, codes aléatoires, HARQ, théorie de l'ergodicité, canal d'effacements, chaîne de Markov.

Acknowledgements

No ship, no matter how strongly built and technologically advanced it is, can make it safely to the destination without the guidance of an experienced captain. Luckily, my PhD-ship had Emre Telatar at its helm. Being a veteran captain, Emre does not content himself in bringing the ship just safely to port but he brings it full of treasures. That's why, I am deeply indebted to Emre for two things: First, for his crucial technical advices that were essential in the development of my technical skills, and without which I would not have completed my work. In fact, I would like to thank him especially for his patience during the first year of my PhD when I was still learning the tools of my trade. Second, I am also grateful for all our non-technical discussions that helped me develop intellectually and acquire a better taste at fine arts. These are treasures that I will always cherish.

I am also grateful to my ex-officemate and my friend Rajai Nasser with whom I spent countless hours trying to solve two kinds of problems: problems related to the subject of this thesis and problems related to the conflicts in the Middle East. Whereas our collaboration was fruitful in solving problems of a technical nature, we are still at lost on how to address even a single conflict in the Levant.

I would also like to thank the members of the jury committee: Anthony Ephremides, Nicolas Macris, Patrick Thiran and Roy Yates. I really appreciate the time they spent on reading this work and their detailed feedback helped improve the final version of this thesis.

During my second year of PhD, I had the chance to visit Rutgers university and work with Emina Soljanin. I would like to use this opportunity to thank her for the wonderful time she helped me spend in the U.S. as well as for the extremely productive collaboration. Here, I would like to extend my thanks to the members of Emina's lab for their welcoming spirit and the friendships that we built: Jing, Fatih, Nadia, Navid, Pei, Chryssalenia. During this stay and thanks to Emina, I had the happy opportunity to meet and work with Roy Yates on multiple problems.

Back in Lausanne, as a member of the Information Processing Group (IPG) at EPFL, I was part of a family that made my journey smooth and the working environment very enjoyable. For that, I would like to thank first the senior members of our group: Bixio Rimoldi, Rüdiger Urbanke, Micheal Gastpar and Olivier Lévêque. In addition to the professors, I would like to thank the invisible soldiers working behind the scenes. A big thanks to Muriel Bardet and Françoise Behn for taking care of all the logistical dimensions during this period. A special thanks also to Muriel for all the administrative advices and for making the lab run without any problems. Another invisible soldier that I would like to thank is Damir Laurenzi, who made

sure we always had the best equipment and who was always there to help whenever some apocalyptic IT problem stroke.

What made the PhD journey enjoyable is the presence of my brothers-in-arms and sister-in-arm with whom I shared many memories in ISIT and in Lausanne. I am grateful to my officemates, Marco, Rajai and Clément, for sharing not just an office with me but also valuable memories and friendships. Thanks also to the other members (current and previous) of IPG: Mani, Young, Eric, Erixhen. A big and special thanks for the Lebanese gang in IPG for all the moral support and help you gave me during these four years: Rafah, Mohammad, Serj, Ibrahim. We have created some memories which became defining moments in my life.

Outside IPG, I had the chance to make many friends from the PhD program with whom I shared many memories and adventures. Special thanks to Renata and Artem for the uncountable fun times we spent together. Also, I am grateful to the members of LTS4 for adopting me as one of their own and for all the lunches, activities and baby-foot matches we shared together. Among the first PhD friends that I made are the first-year residents of the middle room in the PhD-office: Agatha, Ajay, Handan, Mario, Miranda, Utku and Yoannis.

In order for a foreign person to be productive two conditions are necessary: first, to have a nice working environment, and second, to have a rich set of friends who would turn Lausanne into a second home. Since I have already thanked the people responsible for fulfilling the first condition, I would like to express my gratitude to my Lebanese friends in Lausanne and Zurich for all the support and memories they have given me. Special thanks to Abbas, Abdo, Amer, Chris, Elsa, Elio, Farah, Feyiz, John, Hiba, Lama, Marwa, Mireille, Hani & Carole, Raed, Roula, Samer, Sarah, Yamane.

I would also like to extend my thanks to my friends back in Lebanon and in Europe for their continuous support which transcended the thousands of kilometers that separate us. Special thanks to Nancy and Giorgio.

Finally, it is always the hardest to thank the people closest to you. Words become scarce or insufficient and you struggle to fully convey what you really feel with the limited vocabulary at your disposal. I will try my best in the next few lines to express my deep gratitude towards my extended family. A big thanks to my aunts, uncle and cousins who believed in me since day one and fueled me with love and support despite the long distances and the time differences. A very special thanks for my brother Jad for always being besides me whenever I needed him, for being my best friend, and for offering me a second home in Paris whenever I needed a small break from the stress of the PhD life. As for my Mother and my Father, words just come short of encompassing the importance of their support, their endless sacrifices and their worries during all these years. I would have never been able to complete this PhD-journey without them by my side, and for that I will remain happily indebted.

Lausanne, February 8^{th}*, 2019* E.N.

Contents

List of Figures

1 Introduction

Life begins with an act of *communication*. The first action newborns take after only seconds of seeing the light is to announce, through a cry, that they are alive. The parents and doctors present in the room interpret this announcement as a sign that the child is healthy. The limited *alphabet* at the disposal of the newborn ({cry, silence}) renders this act one of the first instances of binary communication. However, life and communication are much more intertwined than that and share a deep relationship. In fact, *communication* is at the origin of *life*. The infants who were just born will grow into men and women with their physical development (at least[1]) dictated by the information stored in their respective DNA sequences. Similar to any storage medium (floppy disk, CDs, USB,...), the DNA *encodes* the genetic information in a format that guarantees an almost negligible modification of its content[2] and that can be accessed at any point in time. Although communication principles govern even the smallest aspects of life, we had to wait until 1948 in order for Claude E. Shannon to provide a systematic mathematical model of generic communication systems in his landmark paper "A Mathematical Theory of Communication" [63]. In this paper, Shannon describes the *communication problem*:

> *The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.*

Fig. 1.1 illustrates the above definition. The receiver's responsibility is to recreate at the destination the message chosen by the information source and transmitted through the channel. The destination and the source does not have to be two different physical entities. They can be the same entity; in which case, the transmission is done in time. In other words, the communication problem becomes a *storage* problem: The

[1] We don't mention the intellectual development since much of the brain's characteristics are still not fully understood hence we would like to avoid any controversial statements.

[2] Some mutations might occur during the cellular division (which can be compared to copying the content of a CD) but the probability of that event occurring is very close to zero.

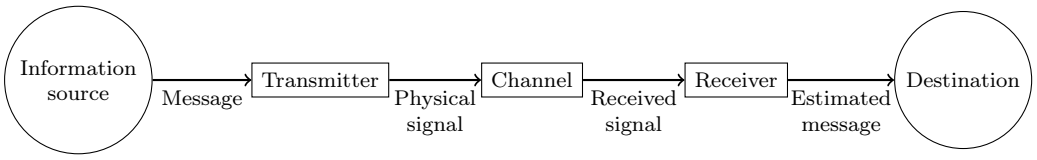


Figure 1.1 – Diagram of a general communication system.

source seeks to access some message it has already generated but at a later moment in the future. This is also known as *data compression* or *source coding*. The storage of the genetic information as DNA sequences falls into this category. Aside from being separated in time, the source and destination could be separated in space. In this case they are necessarily different entities. The newborn crying and two persons talking to each other illustrate this case. In fact, the simple instinctive act of saying "Hello" covers all the components of a communication system as presented in Fig. 1.1: In order for Alice (the information source) to greet Bob (the destination), she first has to choose a message that consists of one or multiple words from the English vocabulary. Having chosen the word "Hello" as her message, she now uses her speaking mechanism (brain, vocal cords, mouth) to convert this message into physical sound waves that propagate through the air to reach the hearing mechanism (brain, ears) of Bob. At this point, Bob's hearing mechanism reconverts the received sound waves into English words that he can understand. In this scenario the speaking mechanism plays the role of the transmitter, the air is the channel, the hearing mechanism is the receiver. In an ideal world, Alice and Bob would be alone and close to each other so that Bob will understand Alice, the first time she speaks. However, it might happen that they are in a public space with much noise around them. This interference might force Alice to repeat her message in order for Bob to correctly understand it. This necessary redundancy comes at the expense of the amount of information Alice can share with Bob in a given interval of time: Every time a message is repeated, Alice loses an opportunity to share a new message. The ratio of the amount of distinct messages over the number of times the channel needs to be used to deliver them is called the *rate* of the communication system. Whereas the process of adding redundancy to a message in order to combat the channel noise is known as *channel coding*.

Until 1948, it was believed that in order to communicate with negligible error probability, the rate of the communication system should be close to zero. In other words, the fewer errors we want to have, the larger the amount of redundancy we need, hence the fewer number of messages we can send. However, Shannon (in his paper [63]) contradicts this belief and shows that every channel $\mathcal{W}$ is characterized by a number $C(\mathcal{W})$, called capacity of the channel; it indicates the highest rate the communication system (using channel $\mathcal{W}$) can adopt while achieving an arbitrarily small probability of error. Any communication system using channel $\mathcal{W}$ that operates at a rate strictly higher than $C(\mathcal{W})$ is guaranteed to have a strictly positive error probability. The rate of a communication system is determined by the channel-coding paradigm adopted by the transmitter. Surprisingly, Shannon proves the existence of a channel code that achieves capacity using a probabilistic approach and thus without suggesting any concrete code. These results paved the way for a plethora of new problems that collectively form the fields of Information Theory and Coding

Theory.

Seven decades after Shannon's work, the technological and theoretical advances[3] in the field of communication and computing opened the door for new categories of applications and services. One such group is real-time status-monitoring systems that are used in healthcare, finance, transportation, smart homes, warehouse and natural environment surveillance, to name but a few. In such systems, a remote monitor is interested in the status of one or multiple processes. A sender takes samples of the observed processes and sends them to the monitor. However, the aim of the communication system in this case is not to transmit as fast as possible but to keep the information the destination has about the observed processes as *fresh* as possible. Let's take the example of Alice following the course of the New York Stock Exchange (NYSE) from her home in Beirut. Although this scenario would have been very unlikely 40 years ago due to the challenges faced in transmitting information over long distances, today we frequently encounter such cases. Therefore, even though the traditional communication problem does not weigh much in the performance of the communication system used by Alice, another question arises: How reliable is the information at Alice's disposal? In fact, Alice's decision to buy or sell a certain amount of company X's stocks is based on the information she received about the status of this enterprise's stock from an observer present at the New York Stock Exchange. According to a pre-chosen algorithm, the observer creates packets at certain instants in time and sends them to Alice. These packets contain the value of company X's stock at the moment of their generation. But, if at a random time t, the information that Alice has about the stock value is "relatively old", the decision she will make will not be optimal and she might even lose money. Hence, although she might receive update packets with negligible error in the content, the information that Alice possesses is still unreliable. Ideally, she would use a communication system that enables her to be constantly up-to-date as if she was physically in the NYSE. However, due to physical constraints such an objective cannot be achieved, but it introduces the engineering problem that we study in this dissertation: How can we design an efficient communication system such that the information the destination has about the status of a remote process is as *fresh* as possible at any point in time? In order to measure the freshness achieved by a system; a new metric called the Age of Information (AoI) is introduced in [33]. Our focus in this thesis is to study this new metric and its behavior under different communication system settings, as well as to compute the lowest achievable age of information for a special category of channels.

The rest of this chapter is organized as follows: In Section 1.1, we give a quick summary of the most important notions in information theory. In Section 1.2, we define in detail the age problem and the age-of-information metric before presenting the most important results in this field in Section 1.3. Finally in Section 1.4, we describe our main contributions of this dissertation.

[3]Technological advances include the development of the transistor, processors, embedded systems...

1.1 The Classic Communication Problem

If two distant parties can communicate, this implies the existence of a physical medium that connects them and through which the communication signal propagates. This medium is called the *channel*.

Definition 1.1. [4] *A channel $\mathcal{W}$ is represented by the triplet $(\mathcal{V}, \mathcal{Z}, Q_{\mathcal{W}})$ where*

- *$\mathcal{V}$ is called the* input alphabet *and each symbol $v \in \mathcal{V}$ is called a* letter. *$\mathcal{V}$ is the set of symbols that are permitted to be sent over the channel. We assume in this text that the input alphabet is discrete.*
- *$\mathcal{Z}$ is called the* output alphabet. *It is the set of symbols that can be observed at the output of the channel. The output alphabet and the input alphabet are not necessarily the same. We also consider only discrete output alphabets.*
- *$Q_{\mathcal{W}}$ is the stochastic matrix that describes the behavior of the channel. This means that $Q_{\mathcal{W}}(z|v) = \mathbb{P}(Z = z|V = v)$ and $\sum_{z\in\mathcal{Z}} Q_{\mathcal{W}}(z|v) = 1$ for all $v \in \mathcal{V}$. We associate the random variable Z with the observed output of the channel and the random variable V with its chosen input. $Q_{\mathcal{W}}(z|v)$ is the probability of observing the letter z at the output of the channel, given that the input letter was v.*

The above definition describes the channel as governed by randomness. For every input letter, $v \in \mathcal{V}$, the channel is characterized by a probability distribution that indicates the relationship between the input letter v and the different output letters. This randomness models the noise that channel might add to the input letter and its statistics are considered to be known beforehand by the transmitter and the receiver. In this dissertation, we are mostly interested in *discrete*, *memoryless* channels. This means that the output of the channel only depends on the current input and not on previous or future ones. One of the most encountered models of discrete memoryless channels is the erasure channel. We will consider this type of channel in Part II of this thesis, that is why we define it here.

Definition 1.2. *An erasure channel $\mathcal{W}$ with erasure probability ϵ is defined by an input alphabet $\mathcal{V} = \{v_1, v_2, \cdots, v_K\}$, with $K \in \mathbb{N}$, an output alphabet $\mathcal{Z} = \mathcal{V} \cup \{?\}$ and for every input letter $v \in \mathcal{V}$ the probability distribution of the output is*

$$Q_{\mathcal{W}}(z|v) = \begin{cases} \epsilon & \text{if } z = ? \\ 1-\epsilon & \text{if } z = v \\ 0 & \text{otherwise.} \end{cases}$$

The above definition implies that if the output of the erasure channel is not $\{?\}$, then we can infer the input from the observation with probability 1. Otherwise, if the output is an erasure $\{?\}$, then the input could be any of the letters in $\mathcal{V}$ with probability $\mathbb{P}(V = v|Z = ?) = \mathbb{P}(V = v)$.

[4] The definitions and results presented in this section are based on the work of Shannon [63].

Example 1.1. *The Binary Erasure Channel (BEC) with erasure probability ϵ is an erasure channel where $\mathcal{V} = \{0, 1\}$ and $\mathcal{Z} = \{0, 1, ?\}$. The probability rule between the input and the output can be represented using the graph in Fig. 1.2.*

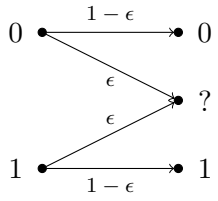


Figure 1.2 – The Binary Erasure Channel (BEC).

In [63], Shannon also uses randomness to model the information source (see Fig. 1.1). The concept of randomness in the source matches well with our intuition: If the output of a source is deterministic and known beforehand, then there is no need to communicate it to a different party.

Definition 1.3. *A source generates symbols $U_1, U_2, U_3, \cdots$ according to a certain random process model. The random variable U_i takes value in the discrete alphabet*[5] *$\mathcal{U} = \{\alpha_1, \alpha_2, \cdots, \alpha_M\}$, where $M \in \mathbb{N}$.*

In this text we are mostly interested in *memoryless* sources; this means that the generated symbols $U_1, U_2, U_3, \cdots$ are independent and identically distributed (i.i.d).

From Fig. 1.1, we gave a formal definition for the information source and the channel. We now elaborate on the transmitter and the receiver.

Definition 1.4. *Given two positive integers k and n,*

- *a transmitter (or an encoder)*[6] *is a function f defined on the set $\mathcal{U}^k = \{u_1 u_2 \cdots u_k;\ u_i \in \mathcal{U},\ \forall\ 1 \leq i \leq k\}$ onto the set $\mathcal{V}^n = \{v_1 v_2 \cdots v_n;\ v_i \in \mathcal{V},\ \forall\ 1 \leq i \leq n\}$:*

$$\begin{aligned} f: \quad & \mathcal{U}^k \to \mathcal{V}^n \\ & u_1 u_2 \cdots u_k \mapsto v_1 v_2 \cdots v_n. \end{aligned}$$

  The encoder associates with each sequence of source letters of length k a sequence of channel input letters of length n. The sequences in $f(\mathcal{U}^k) \subset \mathcal{V}^n$ are called codewords of blocklength n. The set of all the codewords forms a code.

[5] We give the definition for discrete-time discrete-value sources. However, the definition can be easily generalized for continuous time, uncountable alphabet sources.

[6] The definition of encoder given here assumes there is no feedback from the receiver. In the presence of feedback, the definition becomes: Given three positive integers k, n and l, the encoder is a function f such that

$$\begin{aligned} f: \quad & \mathcal{U}^k \times \mathcal{Z}^l \to \mathcal{V}^n \\ & (u_1 u_2 \cdots u_k, z_1 z_2 \cdots z_l) \mapsto v_1 v_2 \cdots v_n. \end{aligned}$$

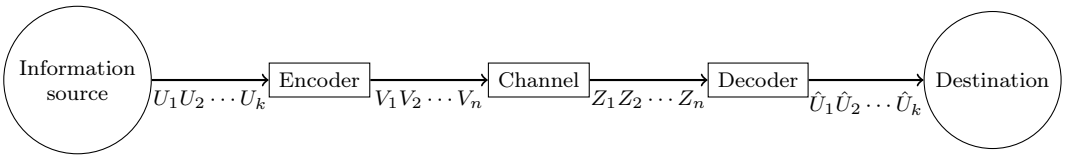


Figure 1.3 – Modified diagram of a general communication system.

- *a receiver (or decoder) is a function g defined on the set $\mathcal{Z}^n = \{z_1z_2\cdots z_n;\ z_i \in \mathcal{Z},\ \forall\ 1 \leq i \leq n\}$ onto the set $\hat{\mathcal{U}}^k = \{\hat{u}_1\hat{u}_2\cdots\hat{u}_k;\ \hat{u}_i \in \hat{\mathcal{U}}^k,\ \forall\ 1 \leq i \leq k\}$:*

$$\begin{aligned} g: \quad & \mathcal{Z}^n \to \hat{\mathcal{U}}^k \\ & z_1z_2\cdots z_n \mapsto \hat{u}_1\hat{u}_2\cdots\hat{u}_k. \end{aligned}$$

  The decoder maps the channel output sequence of length n into a sequence of length k that estimates the sequence generated by the source. To indicate that it is not necessarily equal to the source alphabet, we denote $\hat{\mathcal{U}}$ as the output alphabet of the decoder. However, for all intents and purposes, we assume that $\hat{\mathcal{U}} = \mathcal{U}$ in this thesis.

In this text we adopt the convention where capital letters refer to random variables and lower-case letters refer to deterministic realizations. For instance, V_i corresponds to the random variable associated with the channel input symbol at time i, whereas v is its realization. Similarly, Z_i refers to the channel output symbol at time i, and z refers to its realization. Moreover, the notation V^n will be used to denote a vector of n random variables $(V_1V_2\cdots V_n)$ and v^n to denote a vector of n realization $(v_1v_2\cdots v_n)$.

1.1.1 The Metrics

Based on the aforementioned definitions, the general communication system can be depicted as shown in Fig. 1.3: A source generates in an i.i.d fashion a sequence U^k of symbols that are encoded into a codeword $V^n = f(U^k)$ by the encoder. This codeword is sent one symbol at a time over the channel; and once the decoder observes a sequence Z^n, it tries to retrieve the original source symbols and outputs an estimate $\hat{U}^k = g(Z^n)$. In order to assess the performance of a given communication system, two metrics are used: the communication rate and the block-error probability.

Definition 1.5. *Given a source alphabet $\mathcal{U}$ and an encoder $f : \mathcal{U}^k \to \mathcal{V}^n$, we define the rate of the communication system shown in Fig. 1.3 as*

$$R = \frac{k}{n} \text{ source symbols/channel use.}$$

The rate is the average number of source symbols transmitted per channel use and indicates how fast we can transmit.

We use "log" to refer to the base 2 logarithm and "ln" to refer to the natural logarithm.

Definition 1.6. *Given a source alphabet $\mathcal{U}$, an encoder $f : \mathcal{U}^k \to \mathcal{V}^n$ and a decoder $g : \mathcal{Z}^n \to \mathcal{U}^k$, the block probability of error P_e is the probability that the decoder gives*

an estimate different from the actual transmitted sequence. Formally,

$$P_e = \mathbb{P}(\hat{U}^k \neq U^k) = \mathbb{P}(g(Z^n) \neq U^k).$$

The block error probability measures the reliability of the communication system at hand: The lower the error probability the more reliable our system is.

The goal of a system engineer is to come up, for a given source and channel, with a communication setup (k, n, encoder f, decoder g) that achieves the highest rate possible while ensuring a negligible error probability. This is the classic communication problem.

1.1.2 Source Coding or Noiseless Communication

A noiseless channel $\mathcal{W}$ is a channel where the output is exactly equal to the input. This means that $\mathcal{V} = \mathcal{Z}$ and for all $v \in \mathcal{V}$ we have

$$Q_{\mathcal{W}}(z|v) = \begin{cases} 1 & \text{if } z = v \\ 0 & \text{otherwise.} \end{cases}$$

In his paper [63], Shannon shows that solving the classic communication problem for noiseless channels is equivalent to solving a data compression problem. As the channel in this context is simply a noiseless wire, let's assume that it transmits bits. Thus $\mathcal{V} = \mathcal{Z} = \{0, 1\}$. Hence, the encoder f maps a sequence of k source symbols into a sequence $V^n = (V_1 V_2 \cdots V_n)$ of n bits and the decoder g estimates the original sequence of source symbols using V^n. Moreover, maximizing the communication rate $R = \frac{k}{n}$ (source symbols/bit) is equivalent to minimizing the compression rate $C = \frac{1}{R} = \frac{n}{k}$ (bits/source symbol) that is the average number of bits per source symbol. For a given source generating symbols $(U_l)_{l \geq 1}$ in an i.i.d fashion from an alphabet $\mathcal{U}$, Shannon shows that each source symbol can be compressed losslessly[7] up to

$$\mathrm{H}(U) = -\sum_{u \in \mathcal{U}} \mathbb{P}(U = u) \log\left(\mathbb{P}(U = u)\right) \text{ bits},$$

where U is the generic random variable with the same distribution as the U_l, $l \geq 1$. The quantity H(U) is called the *entropy* of U. It measures, in bits, the amount of randomness that is contained in U.
However, in order to achieve this entropy bound we need to compress a large number of symbols at the same time, i.e. $k \to \infty$ [11, 12]. As we will see later, this is not optimal from an *age* point of view.

1.1.3 Channel Coding or Noisy Communication

In the noisy channel scenario, Shannon assumes that the channel is *discrete* and *memoryless*, meaning

$$\mathbb{P}\left(Z^n = z^n | V^n = v^n\right) = \prod_{i=1}^{n} \mathbb{P}\left(Z_i = z_i | V_i = v_i\right).$$

[7] By losslessly we mean that the decoder g can reconstruct the original sequence correctly from $(V_1 V_2 \cdots V_n)$ with probability 1.

Furthermore, he considers that the source chooses independently at random one message out of M messages that constitutes the source alphabet[8] $\mathcal{U}$. Denoting by U the random variable relative to the chosen message, the encoder maps U to a codeword of blocklength n, $V^n = f(U)$ and sends it over n channel uses to the destination. The decoder uses the observed sequence Z^n and outputs an estimate of the original message, $\hat{U} = g(Z^n)$.

The rate of the communication scheme (also called *coding scheme*) defined by the 4-tuple $(\mathcal{U},\, n,\, f,\, g)$ is

$$R = \frac{\log(M)}{n} \text{ bits/channel use.}$$

Moreover, the block error probability becomes $P_e = \mathbb{P}(\hat{U} \neq U)$. Shannon considers a rate R to be *achievable* if for every $\epsilon, \sigma > 0$ there exists a coding scheme of rate at least $R - \epsilon$ and a block error probability $P_e < \sigma$.

Theorem 1.1. *Given a channel $\mathcal{W}$ with an input alphabet $\mathcal{V}$ and an output alphabet $\mathcal{Z}$, a coding scheme $(\mathcal{U},\, n,\, f,\, g)$ has an achievable rate R if and only if*

$$R \leq C(\mathcal{W}).$$

The quantity $C(\mathcal{W})$ is called the capacity of the channel $\mathcal{W}$ and is defined as:

$$C(\mathcal{W}) = \max_{P_V \in D_{\mathcal{V}}} I(V;Z),$$

where

- $I(V;Z) = \sum_{v \in \mathcal{V},\ z \in \mathcal{Z}} \mathbb{P}\left(V = v,\ Z = z\right) \log\left(\frac{\mathbb{P}(V=v,\ Z=z)}{\mathbb{P}(V=v)\mathbb{P}(Z=z)}\right)$,
- *$D_{\mathcal{V}}$ is the set of all probability distributions P_V of the random variable V on the alphabet $\mathcal{V}$.*

Theorem 1.1 says that, given a channel $\mathcal{W}$, if a coding scheme tries to send at a rate strictly higher than the channel capacity $C(\mathcal{W})$, the block error probability will always be lower bounded by a strictly positive number. Shannon does not give a constructive proof for Theorem 1.1 but uses a random coding argument to show the existence of a coding scheme that achieves channel capacity as the blocklength $n \to \infty$. A similar argument will be used in Chapter 7 to give an upper bound on the minimum achievable *age* over an erasure channel.

1.2 The Age Problem

While the classic communication problem described in Section 1.1 is transmission centric, the age problem presents itself as reception centric. In fact, these two problems study the communication system from two different perspectives: In classic information theory, the transmitter is the main instigator of the transmission process.

[8] Comparing it to the noiseless channel scenario, we assume here that $k = 1$ and we call each source symbol a message.

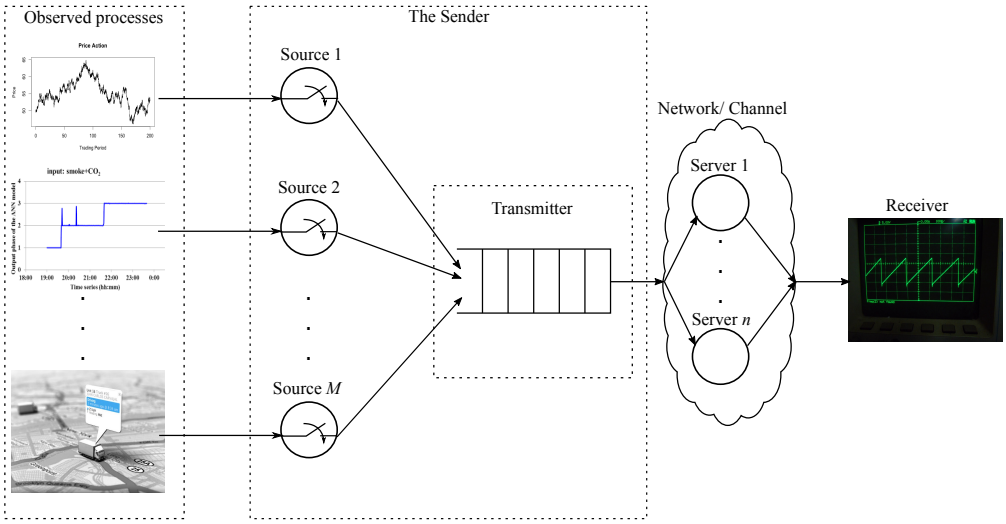


Figure 1.4 – The age problem communication setup.

Indeed, a source wants to share with a *remote* target (*remote* in time or space) the maximum amount of information (usually measured in bits). This can be achieved through transmission at channel capacity that is the maximum possible rate that ensures a negligible decoding error probability. Whereas, in the age of information (AoI) theory, the receiver is the one who initiates the communication. A monitor follows the evolution of the status of one or multiple *remote* processes (in this case *remote* refers only to distance in space) in a timely manner: Specifically, at any time instant t the information the receiver has about the status of a process should be as close as possible to the current status of the process. This goal can be accomplished by adopting the transmission scheme that would minimize a new metric: *the age*. We will see later that there exist different definitions for the *age* metric but first we start by giving a detailed description of the communication frame that is common to all Age of Information (AoI) problems.

1.2.1 The Setup

The communication setup that the age of information literature addresses, shown in Fig. 1.4, can be described as follows:

- *Observed Processes*: These are the set of continuous-time random processes that a distant receiver wants to monitor. This means that the receiver (or interchangeably, the monitor) would like to receive updates describing the status of these processes at multiple instants in time. Examples of such processes include (i) the evolution of the price of a certain commodity in the stock market (process 1 in Fig. 1.4), (ii) the variation of the level of carbon dioxide in a forest (measured using a sensor [19, 39, 42, 57, 60]) to detect fire hazards (process 2 in Fig. 1.4) and (iii) keeping track of the positions of nearby cars in a vehicular network [33, 37, 51] (process M in Fig. 1.4). This last potential application inspired the authors in [33] to introduce the *age* as a new metric.

- *The Sender*: It observes the processes on behalf of the receiver and sends it information on the status of the observed processes, according to some "mechanism". This information is called *status updates* or just *updates*. We will see later that "mechanism" can take on many definitions depending on the generation model assumed. The sender is constituted of two parts:

  - *The Sources*: They have direct access to the processes of interest. A source m observes process m and generates at time t a packet containing the status of the process at this same instant, as well as a time stamp, $u(t) = t$, of the generation time. The time stamp will help the receiver assess how old the information about the status of the observed process is if it receives this packet (see Fig 1.5 for an example of update generated by source M). As the main objective of a packet is to deliver status updates then we will use interchangeably throughout this text the terms *packet*, *status update* and *update*. Each source can adopt its own sampling or generation mechanism that can take the form of a renewal process, a periodic (deterministic) process or just-in-time generation (more on that later). Given that each source represents an observed process and that there are no delays assumed when generating the packets, we can see the sources as the origin of the information, hence the word *source* will hereafter refer to the underlying observed process and its sampling device. Finally, we will denote by *stream* the set of packets produced by a certain source.
  - *The Transmitter/Scheduler*: This is the brain of the sender. It stores the generated packets in one or multiple queues and assigns priorities to the different sources, as well as to the multiple packets generated by the same source. The transmitter decides according to a certain model, for each source, which packets need to be sent and which ones can be discarded. Similarly, at each transmission instant, it chooses with respect to the sources' priorities from which stream to transmit first. Depending on the number of servers the transmitter has at its disposal, it can send one or multiple packets at the same time (each server serves one packet at a time). In the case of noisy channel, the transmitter can encode the packets received from the source to combat the noise. As we will see later, the choices taken by the transmitter play a major role in the computation of the *age*.

  From an engineering point of view, the generation mechanism(s) adopted by the sources, as well as the transmission policy executed by the transmitter, can be seen as design parameters that can be tuned in order to minimize the *age*.

- *Network/Channel*: This is the medium through which the packets are transmitted. It can be the Internet, the air, an erasure point-to-point channel, etc. The channel can be assumed to be noiseless, meaning that from the channel point of view any transmitted packet will be delivered correctly with probability 1. However, we can also consider noisy channels where the content of the packets becomes distorted or the whole packet is simply erased. We will address both models in this work: In Part I we tackle the age problem under the noiseless

Packet	
Status Update	Unit 16 Driving 25 kph Heading NNE
Time-stamp	01.02.2019 8:34 a.m.

Figure 1.5 – A status update packet generated by source M.

channel assumption, whereas in Part II we take a more physical approach and study this problem for the noisy erasure channel model.

- *The Receiver/Monitor*: This is the device that is interested in the status of the remote processes. It poses the condition of having at any point in time the *freshest* information concerning the status of the observed processes. In order to measure this *freshness*, the concept of *average age* was introduced in [33] and defined more rigorously in [34]. This new concept is fundamentally different from the notion of delay, as the latter is packet-centric and measures how long a status update spends in transit and thus does not give the receiver any indication on how stale his knowledge about the remote processes is.

1.2.2 The Age Metrics

We now define the main metrics that quantify the notion of *freshness*. Without loss of generality, we consider a single source. The same concepts can be applied to each one of the multiple sources.

First notice that not all generated packets will necessarily be received by the monitor. In fact, some updates will be discarded by the chosen transmission policies, and others could be erased by the channel. This means that only the successfully received packets will affect the *freshness* of the information at the monitor. We call these packets *successful packets* or *successful updates*.

Definition 1.7. *Denoting by $u(t)$ the generation time of the last successfully received packet before time t, the instantaneous age of the information— relative to the status of a source— at the receiver at time t is defined as*

$$\Delta(t) = t - u(t). \tag{1.1}$$

Note that $\Delta(t)$ is a continuous-time continuous-value stochastic process.

Observe that $\Delta(t)$ increases linearly in the intervals between packet receptions. This agrees with our intuition, because whenever the monitor is waiting for a new update the information it has about the observed process becomes obsolete with each passing time unit. However when the j^{th} successful packet, generated at time t_j, is received at time t'_j, $\Delta(t)$ jumps down to the delay experienced by this packet, namely to $T = t'_j - t_j$. This results in a sawtooth sample path, as shown in Fig. 1.6. Using this

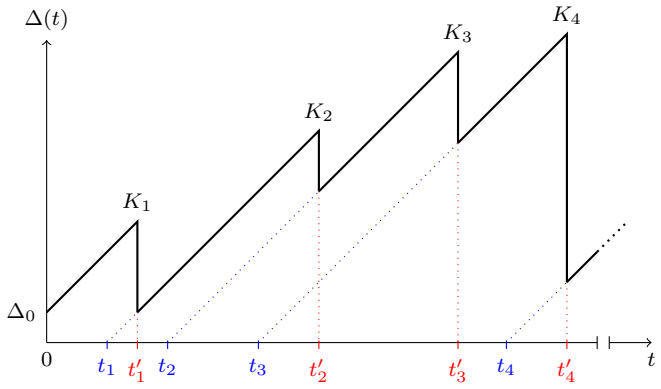


Figure 1.6 – Variation of the instantaneous age for a single source.

definition of the instantaneous age, we are interested in two age metrics: the *average age* and the *average peak age*.

The Average Age

Definition 1.8. *We call Average age the time average of the instantaneous age, given by*

$$\Delta = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^{\tau} \Delta(t) \mathrm{d}t. \tag{1.2}$$

Observe that the average age is the area under the instantaneous age curve shown in Fig 1.6.

This is the first metric that was used to assess, from an age point of view, the performance of a communication system. It was suggested in [33] as a measure of *freshness* in vehicular network then studied in a more theoretical approach in [34]. This last work considers the following setup: One source generates updates according to a Poisson process of rate λ and sends them according to the First-Come-First-Serve (FCFS) policy through the network to the monitor. The transmitter, in this case, is a queue where the generated packets wait for their turn whenever they find the network busy or previous updates also waiting. This network is assumed to be lossless, which means all packets are received correctly at the monitor. However, the time spent by an update in service, i.e., the time spent in the network, is considered to be exponential with rate μ. Such a model is known in Queuing Theory as an M/M/1 queue[9]. Notice that in this model all generated packets are *successful packets*. In this work, the authors make three crucial observations:

[9]This notation is known as Kendall notation. The first letter indicates the distribution of the time interval between the generation of two successive packets. This time interval is call the *interarrival time*. The second letter refers to the distribution of the service time of a packet and the last number points to the number of packets served at a time. In the case of M/M/1 the transmitter can only send one packet at a given time. Moreover, the letter "M" indicates that the considered distribution is the *exponential distribution* that is the only *memoryless* continuous distribution, [58].

1. **The concepts of *age*, *delay*, and *throughput* are fundamentally different**.
   In fact, they might lead to conflicting optimal transmission policies. For instance, assuming a fixed service rate μ, we can increase the *throughput* by increasing the average generation rate λ as much as possible[10]. However, generating updates at an *elevated*[11] rate can lead to congestion, and a high number of packets will be waiting in the queue for their turn to be served. From an age point of view, this implies that the updates are getting "old" while still at the transmitter side. Hence, the average age for systems operating at high average generation rate will be large. Therefore, maximizing the *throughput* and minimizing the *average age* appear to be two contradictory objectives. A similar observation can be made concerning minimizing both *delay* and *average age*. Indeed, from the j^{th} packet perspective, the *delay* time T_j is the interval of time between its generation time t_j and its reception time t'_j, namely $T_j = t'_j - t_j$[12]. In the FCFS model, this *delay* consists of two components: a waiting time and a service time. The waiting time of a packet represents the amount of time this update needed to wait in the queue before starting service. Although the service time is usually dictated by physical realities and constraints, we can still affect the waiting time through manipulation of the update generation process. For example, assuming a fixed average service rate μ, a small average generation rate $\lambda \ll \mu$ means that, on average, a packet service time is much smaller than the time interval between the generation of two successive updates. Therefore, a newly created packet will, on average, find the network free and start its service instantaneously. Hence, the average waiting time will be negligible and the *delay* is dominated by the service time. Whereas $\lambda \ll \mu$ is optimal from a packet *delay* point of view, it is intuitively not the best strategy to adopt from an *age* point of view. This can be explained by noting that while waiting for a new update to be generated, the instantaneous age $\Delta(t)$ at the receiver is increasing linearly with time as the information the monitor has becomes obsolete. Therefore, a small average generation rate λ means a long inter-generation time interval that will negatively affect the *age*.

2. **There exists a non-trivial optimal average generation rate**[13].
   The above observations lead us to suspect that the optimal average generation rate λ^* that minimizes the *average age* should be neither too small ($\lambda \ll \mu$) nor too big ($\lambda \approx \mu$). In fact, it is shown that the optimal operating point is when $\frac{\lambda}{\mu} \approx 0.53$. This means that the average generation rate should be chosen so that the amount of time the network is busy serving packets is slightly longer than the amount of time it is idle.

3. **There exists a non-trivial optimal update generation policy**.
   In [34], the authors also compute the *average age* when assuming a *deterministic* generation mechanism, but they keep the FCFS policy and a service time

[10] In a FCFS M/M/1 queue the average generation rate λ should be strictly less than the average service rate μ. Otherwise, the waiting queue will grow indefinitely and the system becomes unstable.

[11] By *elevated* we mean close to the average service rate μ.

[12] This is also known in the Queuing Theory literature as the *system time.*

[13] We have seen in §1.2.1 that the generation mechanism is a design parameter that we can tune to optimize the average age.

exponentially distributed. This means that packets are created periodically at the exact rate of λ packets per second. In Kendall notation, this model translates into D/M/1 queue. It is shown through simulations that this new update generation strategy performs better than the one discussed before and leads to a lower average age.

The Average Peak Age

The peak age K_j in Fig. 1.6 denotes the instantaneous age just before the reception of the j^{th} successful packet.

Definition 1.9. *Let t_j and t'_j be respectively the generation time and the reception time of the j^{th} successful packet. Then we define the peak age K_j as*

$$K_j = \lim_{\substack{t \to t'_j \\ t < t'_j}} \Delta(t). \tag{1.3}$$

The authors in [9] introduce the *average peak age* as the time average of the peak ages.

Definition 1.10. *The average peak age is given by*

$$\Delta_{peak} = \lim_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} K_j. \tag{1.4}$$

This metric, apart from displaying a similar behavior as the average age, is much more tractable and easier to compute in most cases. In this work, we are interested in both age measures.

Other Functionals of the Age

While the average age and average peak age are considered to be the two main age metrics, one can choose any *measurable*, *non-decreasing*, *non-negative* functional of the instantaneous age as a cost function. This idea is first introduced by Sun et al. in [66] where the authors define the age penalty function $g : [0, \infty) \to [0, \infty)$ to express the level of "dissatisfaction" in data staleness. Thus instead of optimizing the average age or average peak age, the authors aim to minimize the average age penalty defined as follows:

Definition 1.11. *Given a measurable, non-decreasing, non-negative function $g(.)$ and the instantaneous age function $\Delta(t)$, the average age penalty is defined as*

$$\Delta_g = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^{\tau} g(\Delta(t)) \mathrm{d}t. \tag{1.5}$$

Note that if we take g to be the identity function ($g(t) = t$) the average age penalty becomes the average age, $\Delta_g = \Delta$. The notion of age penalty is revisited in [40] where the authors refer to it as *cost of update delay* (CoUD). However, in this work

the choice of the CoUD function $g(.)$ depends on the observed process statistics. On one hand, if the observed process has a high autocorrelation, meaning that any two samples are highly correlated and the status of the process can most likely be inferred using the updates that the receiver already has, then the cost of waiting for a new update at the monitor can be assumed not to increase very fast. This translates into choosing, for example, $g(t) = \log(\alpha t + 1)$, where $\alpha > 0$ is a tuning parameter. On the other hand, if the observed process has a low autocorrelation— meaning that any two samples are weakly correlated and we cannot deduce much of the current status of the process based on previous updates— the cost of information staleness at the monitor becomes important as the uncertainty on the process status quickly increases with time. This idea can be reflected by using an exponential CoUD, $g(t) = e^{\alpha t} - 1$ with $\alpha > 0$, also a tuning parameter. For processes with intermediate autocorrelation the authors in [40] suggest using a linear CoUD, $g(t) = \alpha t$. In addition to the CoUD, Kosta et al. define another measure called the *value of information update* (VoIU).

Definition 1.12. *Given an age penalty function or CoUD $g(.)$ and assuming that the j^{th} update is generated at time t_j and received at time t'_j, then the value of information of update (VoIU) j is*

$$V_j = \frac{g(t'_j - t_{j-1}) - g(t'_j - t_j)}{g(t'_j - t_{j-1})}. \tag{1.6}$$

The VoIU measures the contribution a certain update brings to the reduction of the age penalty function.

The maximum value of V_j is 1 and this is achieved when the j^{th} update is instantaneously received after its generation, thus dropping the instantaneous age to 0. Similarly to the average age penalty, we can define the average value of information.

Definition 1.13. *Assuming that in the interval $(0, \tau)$ the monitor receives $N(\tau)$ updates, then the average value of information is*

$$V = \lim_{\tau \to \infty} \frac{1}{\tau} \sum_{j=1}^{N(\tau)} V_j. \tag{1.7}$$

1.2.3 To Send or not to Send

We already pointed out in §1.2.1 that the choice of the generation and transmission policy greatly affects the value of the *age*. Now that we have defined the metrics used to evaluate the *age*, we still need to define the major scheduling policies. For this purpose, we distinguish between two types of transmission: *Dumb transmission* and *transmission with packet management*. Here too, without loss of generality, we assume a single source.

Dumb Transmission

In this scheme, there is no intelligence at the transmitter side. All generated packets have the same priority and are automatically queued for transmission. No update is dropped or delayed. In this case, we can transfer the decision making to the source instead.

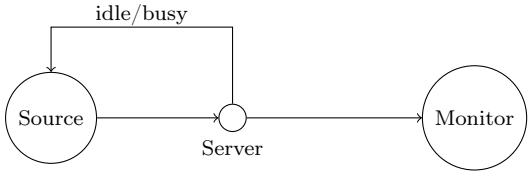


Figure 1.7 – Simplified view of the just-in-time transmission scheme with a single source and single server.

- **First-Come-First-Served (FCFS)**: In this policy, the transmitter consists of one queue with one or multiple servers serving one or multiple packets at the same time. If the scheduler has only a single server, the packets generated by the source are placed in the queue in chronological order with the oldest update at the head. The server starts the transmission of a packet only after the reception of the previous one (we assume an acknowledgment (ACK) is sent back by the monitor upon reception of a new update). If the scheduler has $s > 1$ servers at its disposal, this means that it can transmit up to s packets at the same time. In this case, the transmitter places the newly generated packet in the queue if all the servers are busy. Otherwise, it directs the new update to an idle server. Such schemes assume that the source does not have access to the state of the transmitter (busy or idle). Examples of FCFS policies are G/G/1[14] , G/G/s and G/G/∞[15]. As we have discussed before, [34] deals with M/M/1, D/M/1 and also M/D/1 scheme, whereas [26–28] study the average age for M/M/2 and M/M/∞.

Packet Management at the Source

In the following policies, the transmitter is still primitive but the source has some intelligence.

- **Zero-wait/Just-in-time**: This scheme assumes the source has access to the status of the server(s) and it can generate updates at will. Whenever a server is done transmitting a packet it signals the source which generates instantaneously a new sample (see Fig. 1.7). As the source only generates a packet whenever one of the servers becomes idle, this means that there are no updates waiting in a queue, hence there are no queues. The sender is reduced to just the source and a "dumb" transmitter. Moreover, a new packet starts service immediately upon generation. This policy maximizes the throughput while minimizing the delay per packet. To obtain a lower bound on the average age achieved with a FCFS policy, it was first introduced in [34] then revisited in [66, 71].

- **Lazy generation**: This model revolves around the same concept as the *just-in-time* scheme but with a twist. In addition to the fact that it has access

[14] In Kendall notation, G/G/1 means that the time between the generation of two consecutive updates has a general distribution G. Similarly, the time spent by a packet in service has also a general distribution F (not necessarily the same as G).

[15] In Kendall notation, the s in G/G/s refers to the number of servers or packets that can be sent at the same time. In this case, on can transmit up to s updates at the same time. G/G/1 and G/G/∞ are special cases where $s = 1$ and $s \to \infty$ respectively.

to the status of the server(s) and generates updates at will, the source can delay the generation of a packet if it sees fit (hence the adjective *lazy*). More explicitly, rather than creating an update at the exact moment when the server becomes idle, the source can wait a certain amount of time before generating and submitting its new packet to the transmitter. The source waiting time is a design parameter that can be random or deterministic. Yates in [71] shows that *just-in-time* transmission scheme is not necessarily optimal and that a fine-tuned lazy transmission performs better, an idea further developed in [66, 67].

Transmission with Packet Management

Intuitively, compared to the FCFS policy, giving the transmitter some degrees of freedom and enabling it to choose which packets to send and which ones to discard, we might get better performance from an age point of view. Such a transmitter is said to have packet-management capabilities. There are three main packet-management strategies at the transmitter level that are studied in the age of information literature.

- **Last-Come-First-Served (LCFS) with no buffer or G/G/1/1 with blocking**[16]: In this model the transmitter does not save any packets and consists only of the server. If a new packet finds the server idle, it is immediately served. However, if it finds the server busy transmitting a previous update, the transmitter simply discards the new sample. Hence, there can be at most one packet in the system at a time and all other updates are blocked and dropped. This policy is first introduced in [9, 10] where the authors compute the average age and average peak age for the M/M/1/1 case.

- **G/G/1/2* or Last-Come-First-Served (LCFS) with preemption in waiting**[17]: In this model the transmitter consists of a buffer of size 1 and a server. This means it can store up to 1 packet while serving an additional one. However, if a newly generated update finds the buffer full, the transmitter discards the waiting packet and replaces it with the newcomer. An intuitively less optimal scheme is the G/G/1/2 (without the *) where the transmitter discards any new packets that find the system full. These two strategies were also first studied in [9, 10] for the special cases of M/M/1/2 and M/M/1/2*, where it is shown that the M/M/1/2* outperforms the M/M/1/1 with blocking as well as the M/M/1/2. Such a result agrees with our intuition, because compared to the M/M/1/1 with blocking scheme, the M/M/1/2* increases the throughput by decreasing the amount of time the server is idle waiting for a new update. Furthermore, compared to the M/M/1/2 strategy, the M/M/1/2* decreases the waiting time of the waiting packets as only the most recent generated packet stays in the buffer. These two observations explain the superior performance of the M/M/1/2*. In [52], the authors analyze the behavior of the queue when considering N sources instead of one and an M/M/1/2* policy for each source.

[16] In this notation, the first three entries still have the same meaning as in a classic Kendall notation. The fourth entry, however, refers to the total number of packets allowed in the system at a given point in time (including packets in service).

[17] As for the G/G/1/1 case, the 2 refers to the total number of packets allowed in the system at the same time. In this case, the system can buffer up to 1 packet in addition to the one in service.

- **Last-Come-First-Served (LCFS) with preemption or G/G/1/1 with preemption**: In this model also, the transmitter does not save any packets. However, the priority is given to the newly generated updates. This means that, whenever a new packet is generated and finds the server busy transmitting an old one, the transmitter preempts the transmission of the current update and starts sending the new one. This transmission scheme is proposed in [38] where the authors study the special case M/M/1/1 with preemption.

1.3 AoI and its Applications

1.3.1 Analysis and Optimization

Numerous factors affect the evaluation of the average age and the average peak age, such as the model of the source update process, the number of sources, the model of the transmitter and the number of servers available at the transmitter. The major fraction of the AoI literature tries to answer the following questions: For a given channel model, transmission and source generation policies, what is the optimal update generation rate? More generally, for a given channel model, what is the optimal transmission and/or source generation policies?

Age Analysis in a Single-server Network under an M/M/1/. Scheme

We have already seen that Kaul et al. in [34] solve one aspect of the problem where they consider a single source that generates packets as a rate λ Poisson process feeding them to a single FCFS queue with exponential service time. The authors also consider the cases of a deterministic source and exponential service time, i.e. FCFS $D/M/1$ system, as well as a random source and deterministic service time, i.e. FCFS $M/D/1$ system. Yates and Kaul in [74] generalize the problem solved in [34] by considering the presence of multiple sources that send updates through one FCFS queue to the same monitor. In this case, the goal is to find the region of feasible ages at the receiver and the corresponding optimal update rates at the sources. This leads to the point of the region that minimizes the sum of these ages. As in [34], the queue is assumed to have an exponential service time, and the sources are assumed to be independent Poisson processes with different rates. An interesting result of [74] is that we gain in efficiency if we let two sources share the same queue instead of sending the updates of each source over a separate queue with half the service rate.

In [75], Yates et al. generalize the previous results by assuming an arbitrary number of sources and deriving closed form expressions for the average age when assuming an M/M/1 FCFS policy, an M/M/1/2* and a M/M/1/1 with preemption strategies. In this work, the authors introduce a new analytical method to compute the average age relative to each source, namely the *stochastic hybrid system* (SHS) [21]. In [29, 31], Kam et al. return to the single source single server model but they introduce a new modification on the M/M/1/2 scheme: When a new packet is placed in the buffer of size 1, the transmitter starts a timer. If this timer exceeds a certain deadline and the packet is still waiting in the buffer, the waiting packet is considered too obsolete to be worth transmitting hence is discarded. The authors compute the average age of such model and show that a M/M/1/2 scheme with a carefully chosen deadline

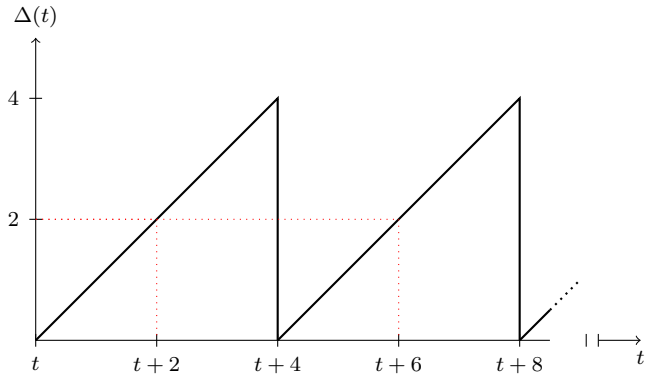


Figure 1.8 – Variation of the instantaneous age for a single just-in-time source.

outperforms the traditional M/M/1/2 as well as the M/M/1/1 with blocking schemes. Kam et al. also study through simulations in [30] the effect of the buffer size, the deadline and packet management on the age. Simulations show that when using packet management schemes (such as M/M/1/2*), enforcing a deadline does not affect the age performance much. Moreover, for the optimal choice of the buffer size and deadline in the case with no packet management, we can potentially get an age performance close to the M/M/1/2* scheme.

Age Analysis in a Single-server with Intelligence at the Source

In [66, 67], Sun et al. replace the random update generation scheme by a generation-at-will model. They show that even if the source is able to create updates at a time instant of its choosing, generating a new packet immediately when the previous one is received (just-in-time generation) is not necessarily the optimal method to adopt. This concept is illustrated by the following example [66, 67].

Example 1.2. *Suppose we have a single source, a single server, and a single monitor. Assuming an idle initial state of the system, the service times of the packets follow this periodic sequence (the unit is seconds): 0,0,2,2,0,0,2,2,0,0,2,2,0,0,... . Let us first consider a just-in-time generation policy and that the first update is generated and delivered at time t. This means that the third generated packet will not give any new information about the source status than the first two updates as they were all generated at time t. Thus we would have wasted our resources to send three times the same packet. As the service time sequence is periodic, the aforementioned problem appears periodically over time: Every time two successive packets have zero service time, we waste the benefit of the second service time by sending the same packet again. The variation of the instantaneous age is given in Fig. 1.8. Intuitively it would be more beneficial after the first packet to wait for a bit before generating the second update so that it would not be redundant. Indeed, from Fig. 1.8 we can see that the average age for the just-in-time model is $\Delta_{just-in-time} = 2$ seconds. However, if we consider a policy which that 0.5 seconds after each update with zero service time before generating a new packet than it can be shown that the average age is $\Delta_{wait} = 1.85$ seconds. Waiting is hence better than the just-in-time strategy.*

This example shows that whenever two successive packets have short service times it is better to introduce a waiting time before generating the second update (lazy generation) so that the latter will have a higher impact on the value of the age. In fact, Sun et al. in [66, 67] show that the just-in-time model is optimal if one of the following conditions is met:

- The correlation between two successive service time random variables is -1.
- The service time process is constant.
- The age penalty function $g(.)$ is constant.

Age Analysis in Multi-server Network

Up till this point in our discussion, the network was assumed to be composed of one hop or server. This means the network is seen as one black box that processes the packet sent by the transmitter according to a certain stochastic model and then delivers it to the monitor. A natural question arises here: How is the age affected if we consider a more granular representation of the network? More explicitly, what is the effect of the presence of multiple hops in series in the network on the age? These multiple hops can be seen as modeling the different routers and switches a packet has to traverse when traveling through the Internet. Bedewy et al. in [6] answer a variant of these questions: Which causal transmission policy at the sender optimizes (in a stochastic order sense[18]) the instantaneous age at the monitor in a multihop network? In this work, it is assumed the packets arrive out-of-order at the transmitter, which has access to a single gateway server, and the service time experienced at each subsequent hop is exponentially distributed. The authors show that the *last-generated-first-served* (LGFS) with preemption transmission scheme at each hop ensures the lowest, in a stochastic order sense, age penalty $g(\Delta(t))$ for any age penalty function $g(.)$. However, no closed-form expression of the average age is given in [6]. Yates et al. in [72] use the SHS method presented in [75] to compute the average age at the receiver, as well as the average ages at each one of the n successive preemptive hops which form the network. Interestingly, this paper shows that even if the hops have different service rates, the order in which they are traversed does not affect the average age at the monitor.

Along the same generalization direction as in [74], we ask: What would happen if we increase the number of servers available at the transmitter, i.e. if the source is able to serve multiple updates at the same time? This question is tackled in [26], where a single Poisson process is sending updates over the network. This can be seen as an infinite number of servers with exponential service rate, $\mathrm{M/M/\infty}$. The challenge here is that packets are received not in a chronological order at the receiver, hence rendering some received packets useless. In this case, the higher the update rate the smaller the average age Δ.

Yates in [73] solve a slight variation of the same problem: The transmitter has access to n servers at the same time and applies a LCFS with preemption policy at each

[18] [61] Let X and Y be two random variables. X is said to be stochastically smaller than Y if $\mathbb{P}(X > x) < \mathbb{P}(Y > x),\ \forall x \in \mathbb{R}$.

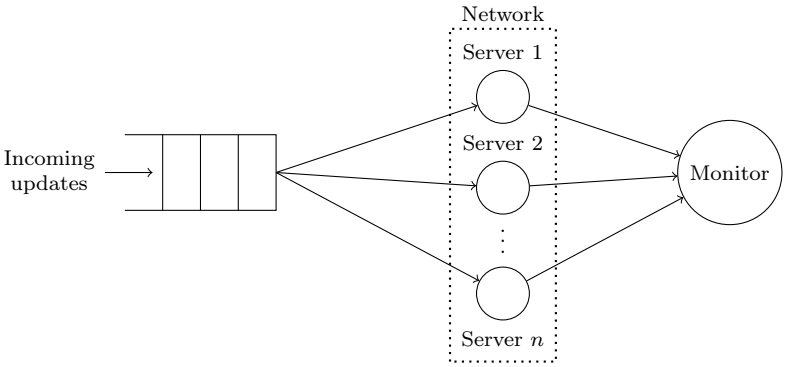


Figure 1.9 – Sketch of the communication setup with multiple/parallel servers. Each server can serve its packets according to its own service time distribution.

server that follows an exponential service time model (see Fig. 1.9). Using SHS, the author derives the expression for the average age and shows that as $n \to \infty$ the average age converges to the formula obtained in [26] for the M/M/∞. The study of the LCFS with preemption policy in such setup is based on the results in [5] where Bedewy et al. show that for a network with n parallel servers (same setup as in Fig. 1.9), the optimal transmission strategy to adopt at each server is the LGFS with preemption. If the updates arrive at the transmitter queue in chronological order, then this scheme is equivalent to the LCFS with preemption. Moreover, if we assume infinite buffer size, Bedewy et al. show that the LGFS with preemption policy optimizes the throughput and the delay.

Age Analysis for Generally Distributed Interarrival Time and/or Service Time

Huang et al. in [22] compute exact expressions of the average peak age when assuming multiple sources sharing a single queue. The novelty stems from the fact that no assumption on the service time distribution relative to each source is made. Two transmission policies are studied: M/G/1 FCFS and M/G/1/1 with blocking (or LCFS with no buffer). Following the same direction of generalization, upper bounds on the average age are given for G/G/1/1 systems (both with preemption and with blocking) in [64]. In a similar line of thought, Inoue et al. in [23] derive the Laplace transform of the stationary distribution of the instantaneous age $\Delta(t)$ and of the peak age when assuming a FCFS G/G/1 scheme. They also compute the average age and average peak age for the special cases of FCFS M/G/1 and FCFS G/M/1 policies. Using results from [22] and [23], Talak et al. in [68] show that deterministic service and interarrival time optimizes the average age and average peak age for FCFS M/G/1 and FCFS G/M/1 respectively. This means

$$\Delta_{FCFS,\ M/G/1} \geq \Delta_{FCFS,\ M/D/1} \qquad \text{and} \qquad \Delta_{FCFS,\ G/M/1} \geq \Delta_{FCFS,\ D/M/1}.$$

Furthermore, they show that for M/G/1/1 with preemption and G/G/∞, deterministic service time leads to the highest average age overall all possible service time distributions, i.e.

$$\Delta_{M/G/1/1,\ preempt} \leq \Delta_{M/D/1/1,\ preempt} \qquad \text{and} \qquad \Delta_{G/G/\infty} \leq \Delta_{G/D/\infty}.$$

An interesting observation pointed in [68] is that for a fixed average service time, the deterministic distribution minimizes the average packet delay when considering the M/G/1/1 with preemption scheme. However, it also maximizes the age. This shows further that the age and the delay are two fundamentally different metrics.

Age Analysis with Source Priority

Up till here, we assumed all sources have the same priorities vis-à-vis transmission precedence. In [36], Kaul et al. rank the sources according to their importance: source 1 has the highest priority and source M has the lowest one. This means that the transmitter gives precedence to packets from source i compared to packets from source $j > i$. Two types of transmission schemes are investigated: (i) an M/M/1/1 with preemption where any new packet from source i preempts the packet currently in service if this update belongs to source j with $j \geq i$, and (ii) an M/M/1/2* where any new packet from source i that finds the server busy would be placed in a buffer of size 1. However, if the buffer is already occupied by an update from source j, $j \geq i$, then the waiting packet is dropped and replaced by the new one from source i. Kaul et al. use the SHS method to compute the average age relative to each one of the M independent sources.

In this thesis, we solve the aforementioned problems under different assumptions, as we will see in Chapters 3 to 5.

1.3.2 Age and Information Theory

In our previous discussion in §1.3.1, we considered mainly the queuing theoretic aspects of the age problem and assumed noiseless channels and perfect reception of packets. However, this approach gives only a high-level understanding of the age-communication challenges without delving into the physical layer problems. Nonetheless, *age* is a communication metric and, as such, is affected by multiple information-theoretic concepts that we have discussed in Section1.1: The source-coding scheme adopted for compression and the channel-coding strategy used to combat the noise in the channel. Whereas the results obtained in the information theory literature mostly assume an asymptotic regime, such an assumption would not be adequate in the age problem. For instance, if we take every packet to be composed of $k \in \mathbb{N}^*$ informative symbols, mapping it to a codeword of length $n \to \infty$ in order to achieve a low probability of error means that the monitor would wait indefinitely for a new update. During this waiting time, the information the monitor has about the source becomes more and more obsolete. Hence, from an age point of view, applying information theoretic results from the finite-blocklength regime (e.g. [54]) appears to be the best approach. Nevertheless, if the blocklength is too small, the probability that the packet will not be correctly decoded and thus will not contribute to decreasing the instantaneous age is high. This means that we might need to send a large number of packets so that one of them is correctly decoded. This too leads to a high average age. Hence we might expect the existence of an optimal blocklength that solves this tradeoff.

Effect of Channel Coding on the Age

Chen et al. in [8] use a model similar to the one used by the works presented before, but they assume that the channel might not deliver the update with a positive probability $1 - p$. This means that the source generates status updates according to a Poisson process with rate λ and sends them through a network with an exponential service time with rate μ. At the end of the service time, the packet might be received correctly with probability $p > 0$, otherwise it is considered lost. Chen et al. compute the average peak age for multiple transmission scenarios: M/M/1 FCFS, M/M/1/1 with preemption, M/M/1/1 with blocking, retransmission with preemption and retransmission without blocking. In the last two schemes, the authors assume an instantaneous lossless feedback that acknowledges the correct reception of the packet. If the packet is declared lost, then the sender retransmits it (instead of just waiting for a new update). The retransmission with preemption scheme is numerically shown to have the best performance.

Parag et al. in [53] consider a more physical approach and assume the channel to be the binary erasure channel (BEC). They also adopt a just-in-time generation process with each packet formed of k information bits and then encoded into an n-bit codeword, $n \geq k$ using a linear code. The authors compute the average age for two transmission schemes: single transmission and hybrid automatic repeat request (HARQ). In the single transmission scheme, every packet is sent only once and after n channel uses a new one is generated and transmitted. Whereas, in the HARQ model, we assume a lossless instantaneous feedback that notifies the transmitter whether the transmitted packet was correctly decoded or not after n channel uses. Moreover, the packet is encoded into αn bits instead of just n bits, with $\alpha \in \mathbb{N}^*$. If the update could not be decoded after the first n channel uses, the transmitter sends the next n bits from the αn-bit codeword. As long as the transmitter receives negative acknowledgment from the receiver it keeps on sending batches of length n bits until the entire codeword is sent. If after sending αn bits the packet could not be decoded, then the transmitter drops it and a new update is generated and starts transmission. For both transmission schemes, the authors show numerically that there exists an optimal value of the blocklength n that minimizes the average age when using random codes.

Yates et al. in [76, 77] also examine the transmission of coded updates generated using a *zero-wait* policy through a binary erasure channel to a monitor/receiver. The authors derive the average status update age of an infinite incremental redundancy (IIR) system in which the transmission of a k-symbol update continues until k symbols are received. This system is then compared to a fixed redundancy (FR) system in which each update is transmitted as an n symbol packet and the packet is successfully received if and only if at least k symbols are received. If fewer than k symbols are received, the update is discarded. Unlike the IIR system, the FR system requires no feedback from the receiver. For a single monitor system, Yates et al. show that tuning the redundancy to the symbol erasure rate enables the FR system to perform almost as well as the IIR system. As the number of monitors is increased, the FR system outperforms the IIR system that guarantees delivery of all updates to all monitors.

The BEC is also treated in [59] where the authors study a FCFS M/G/1 transmission policy. In this case also, the packet at the head of the queue is encoded into a fixed-blocklength codeword before being transmitted. Sac et al. compute the average age and average peak age and show that, in this model also, an optimal blocklength exists. In [14] Devassy et al. explore the additive white Gaussian noise channel (AWGN). The transmitter is considered to follow a FCFS policy with the packet at the head of the queue encoded into a fixed blocklength codeword before transmission. On top of this, the transmitter uses an automatic repeat request (ARQ). This means that, at the end of each packet transmission, the monitor sends a one-bit lossless feedback to the transmitter; this feedback indicates whether the update was correctly decoded or not. As long as the feedback is negative, the transmitter continues to send the same packet until it is correctly decoded. The authors derive the distribution of the peak age and show that there exists an optimal blocklength that minimizes the probability that the peak age is higher than a given threshold.

In this thesis, we focus on the BEC channel and explore different update generation and transmission strategies, as well as different coding techniques (rateless coding in Chapter 6 and random coding in Chapter 7) and their effect on the average age and average peak age.

Effect of Source Coding on the Age

Transmitting maximally compressed symbols (or packets) might appear attractive, at first glance, from an age point of view, because each symbol requires the minimum amount of transmission time. Nevertheless, traditional optimal source coding schemes are in fact not adequate for the age problem. In fact, though optimal source codes reduce the transmission delay to its minimum, they achieve this feat at the expense of a long waiting time. In order for the compression to achieve a rate arbitrarily close to the entropy rate of a stationary source, the number of symbols compressed in one block should be large [11, 12]. This observation leads us to think that the source codes that optimize the compression might not optimize the age. Zhong et al. in [78] address this question by considering fixed-to-variable coding schemes and a FCFS queue. They show through simulations that when the channel service rate is close to the source entropy rate, large blocklength codes should be chosen in order to ensure the stability of the system and that the queue's size does not grow out of bounds. Whereas, in the case where the channel rate is much higher than the source entropy rate, a symbol-by-symbol code achieves the best age performance, because in this regime the average age is mostly dominated by the waiting time needed to form the block that will be compressed. The bigger the block is, the longer the waiting time is. In a subsequent paper [79], the authors assume that the source encoder has access to the state of the channel. They show that, while still using fixed-to-variable codes, dynamically changing the blocklength of the source encoder input, depending on the state of the channel, achieves better performance from an age point of view than adopting a single blocklength size, irrespective of the channel conditions. In [43], Mayekar et al. design a fixed-to-variable prefix-free code that minimizes the average age. The encoder uses a symbol-by-symbol code and if a newly generated symbol finds the channel busy, it is dropped.

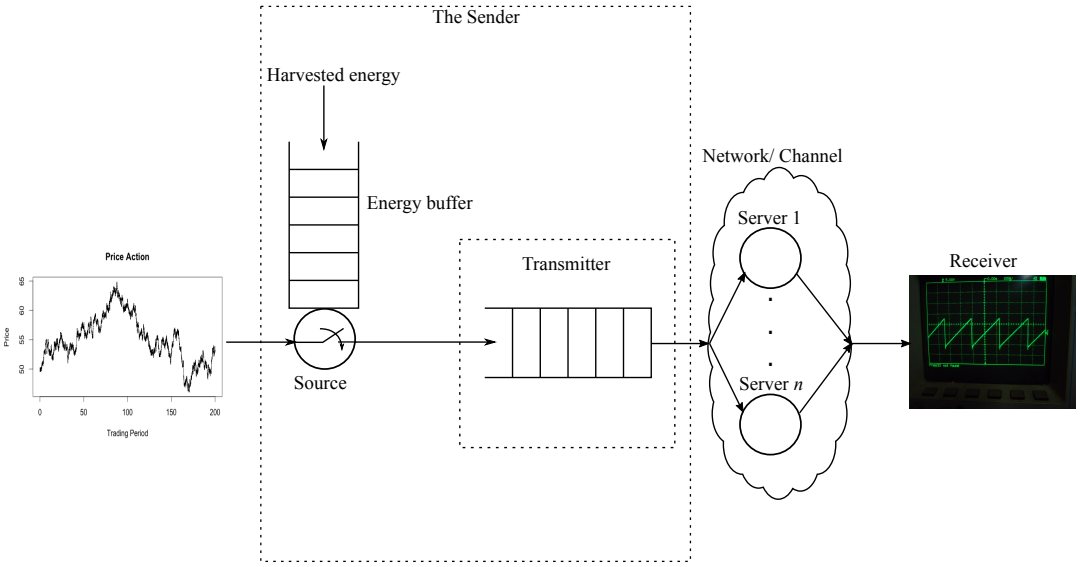


Figure 1.10 – The age problem with energy constraints. The source can only generate an update if there is a sufficient energy amount in the buffer.

1.3.3 Scheduling under Resource Allocation Constraints

More practical models have been studied in the AoI literature, where the scarcity of physical resources is taken into account. The energy constraint is one such resource that is heavily investigated [1–4,15,70,71]. In such models, the number of updates that can be transmitted is upper bounded by an arbitrary time-varying upper bound [2]. This upper bound can be seen as a process (random or deterministic) that mimics the harvest of the energy necessary to the sender to function and transmit the updates. The main setup is shown in Fig 1.10. In [2], Bacinoglu et al. assume instantaneous propagation (service time is zero), which means that no queue is formed. They present the optimal update generation scheme when assuming the energy harvesting process is deterministic and use dynamic programming to solve the problem when the energy harvesting process is stochastic. Yates in [71] considers a just-in-time scheme as well as a lazy generation policies and an i.i.d service process. Moreover, the energy harvesting process is considered to be Poisson. Yates shows that in this context, the just-in-time is not optimal and it is better to wait before generating a new update. The intuition behind this result is that to use the energy to reduce an instantaneous age that is relatively high is better than to reduce an already low age; because the latter scheme might cause us to miss a "better" late opportunity. Other works add different new settings: In [1] Arafa et al. assume the packet is transmitted to the receiver through a relay that is also subject to energy constraints, in [3, 4] Bacinoglu et al. assume a finite battery size and transmission over an erasure channel.

Given that resource allocation is at the core of wireless communications, the primary communication system nowadays, it is natural to study the AoI in such setup. Kadota et al. in [25] study the optimal scheduling policies when assuming a base station sends time-sensitive updates to multiple users. In this case, the metric to optimize is the expected weighted sum of the instantaneous age. In [17], He et al. show that it is NP-hard to find the optimal scheduler that would minimize the peak age when

assuming N sources, N transmitters, and N receivers in the presence of interference. He et al. in [18] generalize this setup by considering M sources and N transmitters. Similarly, they show that the minimum age scheduling problem for this setup is NP-hard. In [24], throughput constraints are considered.

1.3.4 Other AoI Applications

As a measure of freshness, age is a logical metric to use to assess the freshness of cache contents [80]. Moreover, the freshness dimension might suggest that we could use AoI optimal scheduling in order to estimate or predict remote stochastic processes. Sun et al. in [65] show that the sampling policies that optimize the average age does not necessarily minimize the estimation error when the process of interest is a Wiener process. This observation motivated the authors in [32] to suggest two effective age metrics whose optimization implies the minimization of the estimation error: the *sampling age*, which gives the age of the sample with respect to the ideal sampling time, and the *cumulative marginal error*, which represents the total estimation error during a sampling period.

For a detailed survey about the early works in Age of Information, the reader is encouraged to read [41].

1.4 Outline and Main Contributions

This thesis is divided into two parts: Part I "Age in the Absence of Noise", and Part II "Age in the Presence of Noise". In Part I, we compute the average age and average peak age from a queue-theoretic perspective that takes a high-level view on the system and does not make any assumptions on the nature of the physical channel. In Part II, we focus on the erasure channel and shed some light on the behavior of the average age and average peak age over this channel when specific transmission schemes are considered.

However, we first start by presenting, in **Chapter 2**, a quick review of the graphical method used in order to compute the average age and the average peak-age, in the majority of the work on AoI. We also set in this chapter a uniform notation that would be followed throughout this text.

Age in the Absence of Noise

We first start by assuming that the channel consists of a network where packets are delivered with probability 1 but with a random service time. As we have already seen in §1.2.2, the first service time distribution to be studied in the AoI literature was the exponential service time. In **Chapter 3**, we consider a Gamma distributed service time and transmission with packet management. Indeed, in this chapter, we assume that a single source generates updates according to a Poisson process and feeds them to a transmitter that implements either an M/G/1/1 with preemption scheme or an M/G/1/2* scheme. We consider a service time that is gamma distributed for the former policy and that has an Erlang distribution for the latter. Our motivation behind using the gamma distribution is that it is a good approximation for the service

time encountered in networks with multiple relays. We derive closed-form expressions for the average age and the average peak age for both schemes and show numerically that M/G/1/2* has an average age (and average peak age) lower than M/G/1/1 with preemption.

In **Chapter 4**, we generalize part of the results of Chapter 3 in two directions: In the first direction, we consider multiple sources that generate updates according to Poisson processes with different rates and feed them to a single transmitter by applying an M/G/1/1 with preemption policy. Moreover, all sources have the same priority, which means that whenever a packet from source i finds the system busy, the transmitter preempts the update currently in service (irrespective of its source of origin) and transmits the new packet. In the second direction, we assume the same general service time distribution for all packets. The closed-form expressions of the average age and average peak age are derived. For a fixed total update rate, we show that if we want to minimize the aggregate average age, then all sources should be generating updates according to the same rate. However, if we seek to decrease the average age relative to a given source, then we should allocate it a higher generation rate.

In Chapter 4, we assume all sources have equal priority. However, in practical scenarios this is not always the case. Some sources observe processes that are more important to the monitor than others. **Chapter 5** considers this last scenario for two transmission schemes for the low-priority streams: First, we assume the presence of only two sources and a transmitter that adopts an M/M/1/1 with preemption policy for the high-priority source while using a FCFS M/M/1 strategy for the low-priority stream. We compute a closed-form expression for the average peak ages and give an upper and lower bound on the average ages of both streams. Second, we also consider two sources with the difference that both the high-priority and low-priority streams are served by the transmitter according to an M/G/1/1 with preemption strategy. Each stream is assumed to have its own service time distribution (which is considered general). The closed-form expressions for the average ages and average peak ages relative to each stream are derived.

Age in the Presence of Noise

In Part II, channels are not considered ideal anymore and noise is taken into account. To combat such noise different coding schemes are explored and their effect on the average age studied.

In **Chapter 6**, we consider a system where randomly generated updates are to be transmitted to a monitor, but only a single update can be in the transmission service at a time. Therefore, the source has to prioritize between the two possible transmission policies: preempting the current update or discarding the new one. We consider Poisson arrivals and general service time, which means that the two policies we are interested in are the M/G/1/1 with preemption and the M/G/1/1 with blocking. We start by studying the average status-update age and the optimal update-arrival rate for these two schemes under general service time distribution. We then apply these results to two practical scenarios in which updates are sent through an erasure channel by using (a) an infinite incremental redundancy (IIR) HARQ

system and (b) a fixed redundancy (FR) HARQ system. In IIR, the transmission of an update continues until k_s unerased symbols are received. In the FR system, the update is divided into k_p packets encoded ratelessly and each packet is encoded using an (n_s, k_s)-*maximum distance separable* (MDS) code. We show that, in both schemes, the best strategy would be to not preempt. Moreover, we also prove that, from an age point of view, IIR is better than FR.

Chapter 6 raises some interesting questions: Given a single source and an erasure channel, what is the optimal coding scheme from an age point of view and what is the optimal achievable average age? **Chapter 7** answers these two questions in the following two scenarios:

- The source alphabet and the erasure-channel input alphabet are the same. We show that in this case sending the packets without any coding is the optimal policy from an age point of view. Assuming the source generates updates at a deterministic rate of λ updates/seconds and the channel can be used at a deterministic rate of μ channel uses/seconds, we compute a closed-form expression for the optimal average age achieved on an erasure channel.
- The source alphabet and the erasure-channel input alphabet are different. In this case, some kind of encoding is necessary. Assuming the source generates updates at a rate $\lambda = 1$ update/second and we can use the channel at a rate of $\mu = 1$ channel use/second, we use a random coding argument over the linear codes to show the existence of a code that provides an upper bound on the optimal-achievable average age. Using results from [76, 77], we also find a tight approximation of the optimal-achievable average age. Finally, we numerically show that there exists an optimal blocklength that depends on the erasure probability.

System Model and General Settings

2

Kaul et al., in their first analytical paper on the computation of the average age [34], introduced a graphical solution that was later used by the majority of the works in the age of information literature, with various flavors . In this chapter, we present the two main variations of this proof technique and set up a general framework that constitutes the backbone of this thesis.

2.1 General Setup and Notations

We consider the simplified age communication setup shown in Fig. 2.1. In this setup, we assume multiple sources generate updates and send them to the monitor. In this thesis, we adopt the following convention: the average age and average peak-age relative to the status of source i are denoted by Δ_i and $\Delta_{peak,i}$, respectively. Concerning random variables, we use superscript to indicate the source and subscript to indicate the packet or the epoch number. For instance, the random variable $X_k^{(i)}$ is related to source i and to the k^{th} packet of this same source. We show later that an important cornerstone in the derivation of the average age and average peak age is the ergodicity of the status updating system. This means that we deal with steady-state versions of the different random variables involved in these computations. Such variables are denoted by dropping the subscript index from the non steady-state version. For instance, $X^{(i)}$ refers to the steady-state version of the variable $X_k^{(i)}$ relative to stream (or source) i. Whenever it is clear from context which source we are interested in, we drop the superscript notation.

In [75], Yates et al. present a generic graphical method for computing the average age relative to source i without any assumptions on the adopted policy or the probability distributions of the different random variables. The only hypothesis needed is that the age process $\Delta(t)$ is ergodic. For completeness, we repeat this argument here and use this opportunity to define the random variables that we will encounter regularly in the following chapters. We present the derivation for source i only because the

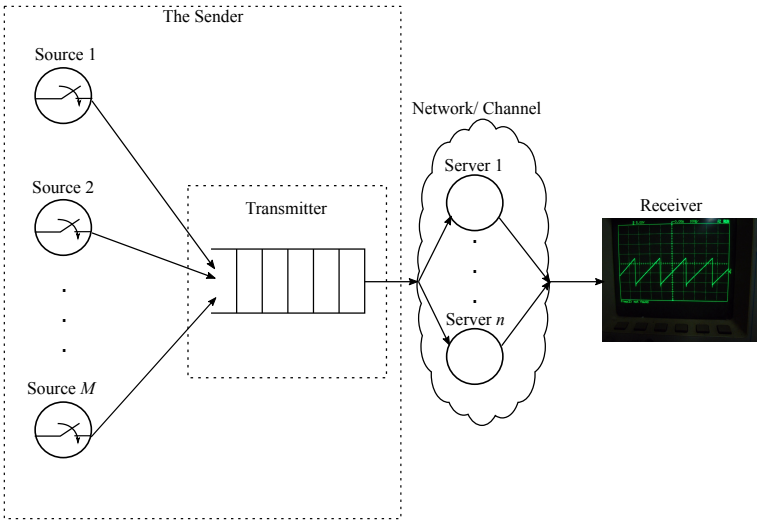


Figure 2.1 – The simplified age communication setup.

same argument and the same quantities can be applied for all other sources. Hence, unless stated otherwise, all quantities in the remainder of this chapter are related to source i (we drop the superscript notation).

Using the model presented in Section 1.2 and in Fig. 2.1, we assume the sender chooses a status update policy Π for source i. This policy defines the updates' generation and transmission schemes. As mentioned before in §1.2.3, some transmission schemes (e.g. M/G/1/1 with preemption or M/G/1/1 with blocking) discard some updates, which means that not every generated packet is necessarily received. In order to differentiate between these two types of updates we use *successful updates/packets* or *delivered updates/packets* to refer to source i updates that were successfully received by the monitor. We denote by I_j the number of packets generated up to and including the j^{th} successful update or

$$I_j = \min\left\{k \geq 1; \sum_{l=1}^{k} \mathbb{1}_{\left\{l^{th} \text{ generated packet is received}\right\}} = j\right\}, \tag{2.1}$$

where $\mathbb{1}_{\{.\}}$ is the indicator function.

Fig. 2.2 shows a snapshot of the instantaneous age relative to source i. Without loss of generality we assume this snapshot was taken at time $t = 0$ when the age value is Δ_0. This plot shows the generation and reception times of the successful packets: The j^{th} *successful update* is generated at time t_j and received at time t'_j. While the monitor is waiting for a new *delivered packet* from source i, the age relative to this source increases linearly. The computation of the age metrics involve different quantities which we now define, and we fix the notation to be used throughout this dissertation.

2.1.1 Interarrival Time

We assume that, according to Π, the time intervals between the generation of two consecutive updates (not necessarily successful) are given by the random process

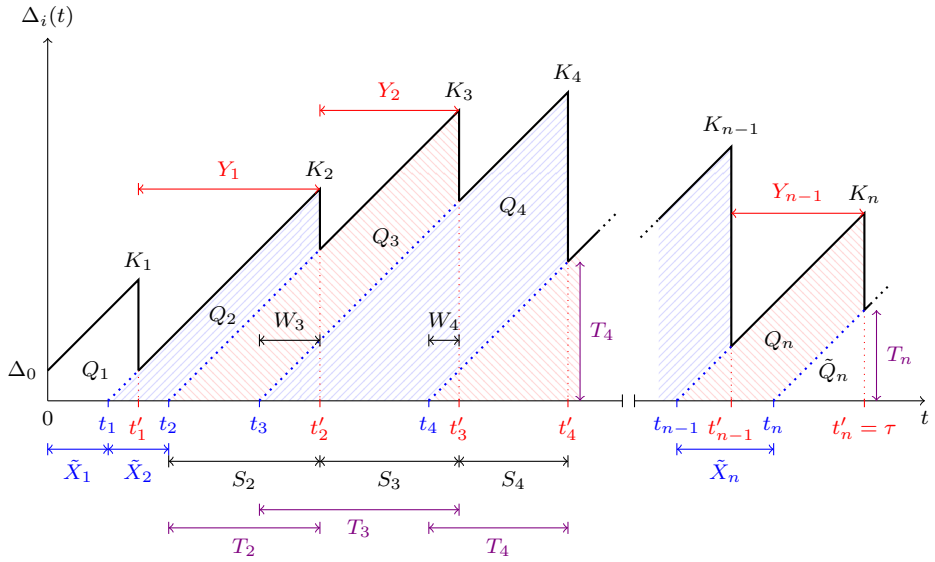


Figure 2.2 – Variation of the instantaneous age for source i.

$(X_k)_{k\geq 1}$. The random variable X_k is called the *interarrival time* between the $k-1^{th}$ and k^{th} consecutively generated packets. However, as only successful packets affect the age $\Delta_i(t)$, we also define the random process $(\tilde{X}_j)_{j\geq 1}$ where the random variable $\tilde{X}_j$ denotes the time interval between the generation of the $j-1^{th}$ and j^{th} delivered updates. We call $\tilde{X}_j$ the *effective interarrival time* and it can be written as

$$\tilde{X}_j = t_j - t_{j-1} = \sum_{k=I_{j-1}+1}^{I_j} X_k, \tag{2.2}$$

where I_j is given by (2.1) and t_j and t_{j-1} are the generation time of the $j-1^{th}$ and j^{th} successful packets (see Fig. 2.2).

Example 2.1. *A generation scheme that is recurrent in the AoI literature assumes the interarrival times X_k to be independent and identically distributed (i.i.d). This means that the number of* events *or* renewals *(in this case the number of generated updates) in an interval $[0,t)$ forms a renewal process [16, 58]. A particular instance of renewal processes is the Poisson process. A Poisson process $N(t)$ of rate λ is a renewal process with the interarrival times being distributed according to an exponential distribution with rate λ. Denoting by X a random variable with such distribution, the probability density function (pdf) of X, $f_X(x)$, is*

$$f_X(x) = \lambda e^{-\lambda x}, \qquad \text{for } x \geq 0.$$

Moreover, the probability that the number of renewals $N(t)$ in the interval $[0,t)$ is equal to $k \geq 0$ is given by the Poisson probability mass function (pmf)

$$\mathbb{P}\left(N(t)=k\right) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

The Poisson process is used in the queuing theory literature to model packets' generation or arrivals *and will be adopted in Chapters 3 to 6. One attractive feature of the*

exponential distribution is its memoryless *property given by*

$$\mathbb{P}\left(X > x + s | X > x\right) = \mathbb{P}\left(X > s\right) \qquad \forall\ x, s > 0.$$

From a packet generation point of view, the memoryless property translates into the following: Given that we already waited x seconds since the last arrival (or generation), the probability distribution over the next s seconds is the same as the probability distribution over the first s seconds. This means that, at any moment in time, the exponential random variable X forgets its history and stochastically resets as if it were starting fresh.

Example 2.2. *If the status update policy $\Pi =$ FCFS M/M/1 queue, all generated packets are received. In this case, the interarrival time process $(X_k)_{k\geq 1}$ is the same as the effective interarrival time process $(\tilde{X}_j)_{j\geq 1}$ and they are an i.i.d process with each X_k (or $\tilde{X}_j$) distributed according to the exponential distribution of rate λ.*

Example 2.3. *If $\Pi =$ M/G/1/1 with preemption policy, we consider that the transmitter has access to a single server with no buffer to store packets that are not in service. This means that any update that finds the system busy is discarded. Hence, not all generated packets are successful and the effective interarrival process $\tilde{X}_j$ differs from the interarrival process $(X_k)_{k\geq 1}$. However, the "M" in M/G/1/1 concerns the interarrival time and not the effective interarrival time. Thus $(X_k)_{k\geq 1}$ is an i.i.d exponential process.*

2.1.2 Service Time

Depending on the policy Π, some packets might be transmitted through the network while some packets might be dropped even before transmission. For the transmitted updates (they could be successful or not), we define the *system time* as the interval of time spent by the transmitted packets in the network. For a certain transmitted packet k we denote its service time by S_k. In all the policies that we discuss in this thesis, we assume the interarrival-time process $(X_k)_{k\geq 1}$ and the service-time process $(S_k)_{k\geq 1}$ to be independent.
In Fig. 2.2, the second successful packet generated at t_2 finds the system empty and thus is instantaneously transmitted. As it is received at t'_2, this means that its service time $S_2 = t'_2 - t_2$. On the contrary, the third successful packet generated at t_3 had to wait for the previous update to leave the network before being transmitted. Thus in this case, its service time is given by $S_3 = t'_3 - t'_2$. If the transmitter can only transmit one packet at a time, the service time of the j^{th} successful packet is given by

$$S_j = t'_j - \max(t'_{j-1}, t_j). \tag{2.3}$$

Example 2.4. *If $\Pi =$ FCFS M/M/1 queue, the service time process $(S_k)_{k\geq 1}$ is an i.i.d process with each S_k distributed according to the same exponential distribution of rate μ. The interarrival time and service processes, respectively $(X_k)_{k\geq 1}$ and $(S_k)_{k\geq 1}$, are considered to be independent from each other.*

Example 2.5. *If $\Pi =$ M/G/1/1 with preemption policy, all generated packets are transmitted. Moreover, the service-time process $(S_k)_{k\geq 1}$ is considered to be i.i.d with each random variable distributed according to a general probability density function*

$f_S(t)$. In this policy also, the interarrival-time process and the service-time process are independent.

2.1.3 Waiting Time

Some policies Π assume the transmitter has a buffer and stores a number of packets when all the servers at its disposal are busy. In this case, the packets in the buffer have to wait for their transmission turn to come hence incur a *waiting time*. The only waiting times that are relevant from an age point of view are those experienced by the successful packets. For the j^{th} successful packet, we denote by W_j its waiting time. From Fig. 2.2, we notice that the second successful packet finds the system empty hence its waiting time $W_2 = 0$. Whereas the forth successful packet finds the system busy serving the third successful update, hence it has to wait for $W_4 = t'_3 - t_4$. If the transmitter can transmit only one packet at a time, the waiting time of the j^{th} successful packet is given by

$$W_j = \max(0, t'_{j-1} - t_j). \tag{2.4}$$

2.1.4 System Time

For any policy Π, the *system time* concerns only successful packets. We define the system time T_j of the j^{th} successful update as the time elapsed between the generation and the reception by the monitor of this update. In other words, T_j is the time spent by the j^{th} successful update in the system (waiting and in service) and it is given by

$$T_j = t'_j - t_j = W_j + S_j. \tag{2.5}$$

In Fig. 2.2, we see that $T_2 = S_2$ but that $T_3 = W_3 + S_3$.

2.1.5 Interdeparture Time

The interdeparture[1] time Y_j is the interval of time elapsed between the reception of the j^{th} and the $j + 1^{th}$ successful updates. This means that

$$Y_j = t'_{j+1} - t'_j. \tag{2.6}$$

2.2 The InterArrival Time Approach (ATA)

Now that the important quantities are defined, we can present the first method for computing the average age and the average peak age: The interArrival Time Approach (ATA). We assume an arbitrary status-updating policy Π, which leads to the age snapshot seen in Fig. 2.2 for source i. As we have already mentioned in the previous section, the sawtooth shape is due to the fact that whenever the monitor waits for a new update, the age of the information on the status of source i, $\Delta_i(t)$ increases linearly with time. However, when the j^{th} successful update is received, the instantaneous age drops to the current age of this last received packet, given by its system time $T_j = t'_j - t_j$. Furthermore, we assume the snapshot covers the interval

[1]The name "interdeparture" is borrowed from queuing theory where delivered packets are said to *depart* from the queue while newly generated packets are said to *arrive* at the queue.

$[0, \tau]$; and without loss of generality, we consider $\tau = t'_n$, the reception time of the n^{th} successful packet. We denote by $N_i(\tau)$ the number of successful source i packets generated up to time $t = \tau$, $N_i(\tau) = \sup\{j \in \mathbb{N}; t_j \leq \tau\}$. In this particular example, $N_i(\tau) = n$.

2.2.1 Computing the Average Age (AoI)

Let

$$\Delta_{\tau,i} = \frac{1}{\tau} \int_0^\tau \Delta_i(t) \mathrm{d}t. \tag{2.7}$$

$\Delta_{\tau,i}$ is equal to the normalized area under the curve $\Delta_i(t)$, for $t \in [0, \tau]$. From Fig. 2.2, the reader can notice that this area could be divided into $N_i(\tau) + 1$ geometric parts: The polygon with area Q_1, the isosceles triangle of area $\tilde{Q}_{N_i(\tau)}$ and the $N_i(\tau) - 1$ trapezoids of areas Q_j, $j = 2, \cdots, N_i(\tau)$. Thus, $\Delta_{\tau,i}$ can be rewritten as

$$\begin{aligned}
\Delta_{\tau,i} &= \frac{1}{\tau} \left(Q_1 + \tilde{Q}_{N_i(\tau)} + \sum_{j=2}^{N_i(\tau)} Q_j \right) \\
&= \frac{Q_1 + \tilde{Q}_{N_i(\tau)}}{\tau} + \frac{N_i(\tau) - 1}{\tau} \frac{1}{N_i(\tau) - 1} \sum_{j=2}^{N_i(\tau)} Q_j.
\end{aligned} \tag{2.8}$$

Recall from Chapter 1 that the average age relative to source i is given by

$$\Delta_i = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau \Delta_i(t) \mathrm{d}t.$$

This means that

$$\begin{aligned}
\Delta_i &= \lim_{\tau \to \infty} \Delta_{\tau,i} \\
&= \lim_{\tau \to \infty} \left[\frac{Q_1 + \tilde{Q}_{N_i(\tau)}}{\tau} + \frac{N_i(\tau) - 1}{\tau} \frac{1}{N_i(\tau) - 1} \sum_{j=2}^{N_i(\tau)} Q_j \right] \\
&= \lim_{\tau \to \infty} \frac{N_i(\tau) - 1}{\tau} \frac{1}{N_i(\tau) - 1} \sum_{j=2}^{N_i(\tau)} Q_j,
\end{aligned} \tag{2.9}$$

where the third equality is due to the fact that the areas Q_1 and $\tilde{Q}_{N_i(\tau)} = \frac{T_n^2}{2}$ form boundary effects hence they are finite with probability 1. This implies $\lim_{\tau \to \infty} \frac{Q_1 + \tilde{Q}_{N_i(\tau)}}{\tau} = 0$.

The ATA consists in writing the trapezoidal area Q_j (hashed surfaces in Fig. 2.2) as the difference of the areas of two right isosceles triangles: One big triangle with a side of length $\tilde{X}_j + T_j$ and one small triangle with a side of length T_j. Thus,

$$\begin{aligned}
Q_j &= \frac{\left(\tilde{X}_j + T_j\right)^2}{2} - \frac{T_j^2}{2} \\
&= \frac{\tilde{X}_j^2}{2} + \tilde{X}_j T_j.
\end{aligned} \tag{2.10}$$

Before going any further, we need to give the following definitions:

Definition 2.1. *Let $(Z_n)_{n\in\mathbb{Z}}$ be a stationary discrete-time stochastic process with $\mathbb{E}(Z) < \infty$, where Z is a generic random variable with the same distribution as any of the Z_n, $n \in \mathbb{Z}$. The process $(Z_n)_{n\in\mathbb{Z}}$ is said to be* mean-ergodic *if the time average of the first moment converge to its ensemble average, i.e.,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Z_i = \mathbb{E}(Z), \tag{2.11}$$

with probability 1.

Definition 2.2. *Let $(Z_n)_{n\in\mathbb{Z}}$ be a stationary discrete-time stochastic process with $\mathbb{E}(Z) < \infty$ and $\mathbb{E}(Z^2) < \infty$, where Z is a generic random variable with the same distribution as any of the Z_n, $n \in \mathbb{Z}$. The process $(Z_n)_{n\in\mathbb{Z}}$ is said to be* second-order-ergodic *if the time average of the first and second moments converge to their ensemble averages, meaning that with probability 1*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Z_i = \mathbb{E}(Z) \qquad \textit{and} \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Z_i Z_{i+k} = \mathbb{E}(Z_0 Z_k) \ \forall\ k \in \mathbb{Z}. \tag{2.12}$$

Definition 2.3. *Let $(Z_n, R_n)_{n\in\mathbb{Z}}$ be a stationary discrete-time stochastic process with marginal distributions similar to (Z, R). This means that the processes $(Z_n)_{n\in\mathbb{Z}}$ and $(R_n)_{n\in\mathbb{Z}}$ are both stationary and have marginal distributions identical to Z and R respectively.The process $(Z_n, R_n)_{n\in\mathbb{Z}}$ is said to be* jointly second-order-ergodic *if with probability 1*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Z_i = \mathbb{E}(Z), \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} R_i = \mathbb{E}(R), \tag{2.13}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Z_i Z_{i+k} = \mathbb{E}(Z_0 Z_k), \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} R_i R_{i+k} = \mathbb{E}(R_0 R_k),\ \forall\ k \geq 0 \tag{2.14}$$

$$\textit{and} \ \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Z_i R_i = \mathbb{E}(XT). \tag{2.15}$$

We now state the main theorem of this section.

Theorem 2.1. *Assume that the status updating policy Π is such that the process $(\tilde{X}_j, T_j)_{j\geq 1}$ is stationary jointly second-order-ergodic with marginal distributions similar to $(\tilde{X}, T)$. In particular, this means that*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i = \mathbb{E}(\tilde{X}), \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} T_i = \mathbb{E}(T),$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i^2 = \mathbb{E}(\tilde{X}^2), \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} T_i^2 = \mathbb{E}(T^2)$$

$$\textit{and} \ \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i T_i = \mathbb{E}(XT).$$

For such an updating policy Π*, the process* $(Q_j)_{j\geq 2}$ *is stationary and mean-ergodic with a marginal distribution similar to* $Q = \frac{\tilde{X}^2}{2} + \tilde{X}T$*, and the average age (AoI) relative to source* i *is*

$$\Delta_i = \frac{\mathbb{E}(Q)}{\mathbb{E}(\tilde{X})} = \frac{\mathbb{E}\left(\tilde{X}_j^2\right) + 2\mathbb{E}\left(\tilde{X}_j T_j\right)}{2\mathbb{E}(\tilde{X})}. \tag{2.16}$$

Remark 2.1. *In [75], Yates et al. refer to a status-updating system with stationary, jointly second-order-ergodic* $(\tilde{X}_j, T_j)_{j\geq 1}$ *process as a stationary and ergodic system. We take a slightly different approach as we consider the term* ergodic *to infer an implication much stronger than the convergence, to their ensemble averages, of the time average of the first and second moments of a process.*

To prove Theorem 2.1 we first need the following lemma:

Lemma 2.1. *let* $N(t)$ *be a counting process. This means that at instant* $t > 0$*,* $N(t) \in \mathbb{N}$ *is the number of events occurring in the interval* $[0, t]$*. Let* Z_i *be the random variable representing the time interval between the occurrence of the* $i - 1^{th}$ *and* i^{th} *events. If the process* $(Z_i)_{i\geq 1}$ *is stationary second-order-ergodic*[2]*, the following equality holds*

$$\lim_{t\to\infty} \frac{N(t)}{t} = \frac{1}{\mathbb{E}(Z)}. \tag{2.17}$$

Proof. By definition, $N(t) = \sup\{n : \sum_{i=1}^{n} Z_i \leq t\}$. This means that

$$\sum_{i=1}^{N(t)} Z_i \leq t \leq \sum_{i=1}^{N(t)+1} Z_i.$$

Thus,

$$\frac{1}{N(t)} \sum_{i=1}^{N(t)} Z_i \leq \frac{t}{N(t)} \leq \frac{N(t)+1}{N(t)} \frac{1}{N(t)+1} \sum_{i=1}^{N(t)+1} Z_i.$$

As $t \to \infty$, $N(t) \to \infty$ and $\frac{N(t)+1}{N(t)} \to 1$. Moreover, since $(Z_i)_{i\geq 1}$ is stationary second-order-ergodic, then $\lim\limits_{N(t)\to\infty} \frac{1}{N(t)} \sum\limits_{i=1}^{N(t)} Z_i = \lim\limits_{N(t)\to\infty} \frac{1}{N(t)+1} \sum\limits_{i=1}^{N(t)+1} Z_i = \mathbb{E}(Z)$. This proves our claim. □

We now provide the proof for Theorem 2.1.

Proof of Theorem 2.1. Let $(\tilde{X}_j, T_j)_{j\geq 1}$ be a stationary jointly second-order-ergodic process as in Definition 2.3. This means that each of the processes $(\tilde{X}_j)_{j\geq 1}$ and $(T_j)_{j\geq 1}$ is stationary second-order-ergodic.

Since for any $\tau > 0$ $N_i(\tau) = \sup\{j \in \mathbb{N} : \sum\limits_{k=1}^{j} \tilde{X}_k \leq \tau\}$, then using Lemma 2.1 we get

$$\lim_{\tau\to\infty} \frac{N_i(\tau) - 1}{\tau} = \frac{1}{\mathbb{E}(\tilde{X})}.$$

[2]For Lemma 2.1 it is sufficient to have a stationary mean-ergodic process.

From (2.10) we can deduce that the process $(Q_j)_{j\geq 2}$ is stationary and mean-ergodic with a marginal distribution similar to $Q = \frac{\tilde{X}^2}{2} + \tilde{X}T$, as it is a quadratic function of a stationary jointly second-order-ergodic process. Thus,

$$\begin{aligned}\lim_{N_i(\tau)\to\infty} \frac{1}{N_i(\tau)-1} \sum_{j=2}^{N_i(\tau)} Q_j &= \lim_{N_i(\tau)\to\infty} \frac{1}{N_i(\tau)-1} \sum_{j=2}^{N_i(\tau)} \frac{\tilde{X}_j^2}{2} + \tilde{X}_j T_j \\ &= \frac{\mathbb{E}\left(\tilde{X}^2\right)}{2} + \mathbb{E}\left(XT\right) \\ &= \mathbb{E}(Q)\end{aligned}$$

where the second equality is justified by the fact that $(\tilde{X}_j, T_j)_{j\geq 1}$ is stationary jointly second-order-ergodic. Hence

$$\Delta_i = \frac{\mathbb{E}(Q)}{\mathbb{E}(\tilde{X})}.$$

Replacing $\mathbb{E}(Q)$ by its expression, we obtain the second equality of (2.16). □

2.2.2 Computing the Average Peak Age (PAoI)

Theorem 2.2. *Assume that the status updating policy* Π *is such that the process* $(\tilde{X}_j, T_j)_{j\geq 1}$ *is stationary jointly second-order-ergodic with marginal distributions similar to* $(\tilde{X}, T)$*. Then the average peak age (PAoI) relative to source* i *is*

$$\Delta_{peak,i} = \mathbb{E}(\tilde{X}) + \mathbb{E}(T) \tag{2.18}$$

Proof. From (1.4), we know that

$$\Delta_{peak,i} = \lim_{N_i(\tau)\to\infty} \frac{1}{N_i(\tau)} \sum_{j=1}^{N_i(\tau)} K_j.$$

From Fig. 2.2 we observe that $K_j = \tilde{X}_j + T_j$, for all $j \geq 1$. As we are assuming the status updating policy Π to be such that $(\tilde{X}_j, T_j)_{j\geq 1}$ is stationary jointly second-order-ergodic, then $(K_j)_{j\geq 1}$ is a stationary mean-ergodic process with a marginal distribution similar to $K = \tilde{X} + T$. Thus,

$$\begin{aligned}\Delta_{peak,i} &= \lim_{N_i(\tau)\to\infty} \frac{1}{N_i(\tau)} \sum_{j=1}^{N_i(\tau)} K_j \\ &= \mathbb{E}(K) \\ &= \mathbb{E}(\tilde{X} + T) \\ &= \mathbb{E}(\tilde{X}) + \mathbb{E}(T)\end{aligned}$$

□

2.3 The InterDeparture Time Approach (DTA)

As in ATA, the trapezoidal area Q_j is also expressed as the difference of the areas of a two right isosceles triangles. Whereas in the ATA we choose to express the length of the side of the big triangle as the sum of the effective interarrival time and system time $(\tilde{X}_j + T_j)$, in the DTA we express it as the sum of the interdeparture time and the system time $(T_{j-1} + Y_{j-1})$. The side of the small triangle is still of length T_j. Thus

$$Q_j = \frac{(T_{j-1} + Y_{j-1})^2}{2} - \frac{T_j^2}{2}. \tag{2.19}$$

2.3.1 Computing the Average Age

Theorem 2.3. *Assume that the status updating policy* Π *is such that the process* $(Y_j, T_j)_{j\geq 1}$ *is stationary jointly second-order-ergodic with marginal distributions similar to* (Y, T)*. For such a policy, the process* $(Q_j)_{j\geq 2}$ *is stationary and mean-ergodic with a marginal distribution identical to* $Q = \frac{Y^2}{2} + YT$*. Then the average age (PAoI) relative to source* i *is*

$$\Delta_i = \frac{\mathbb{E}(Q)}{\mathbb{E}(Y)} = \frac{\mathbb{E}\left(Y^2\right) + 2\mathbb{E}(YT)}{2\mathbb{E}(Y)}. \tag{2.20}$$

Proof. Let $(Y_j, T_j)_{j\geq 1}$ be a stationary jointly second-order-ergodic process with marginal distributions similar to (Y, T). Denote by $R_i(\tau)$ the number of successful receptions of source i packets up to time $t = \tau$, i.e. $R_i(\tau) = \sup\{j \in \mathbb{N} : t'_j \leq \tau\}$, where t'_j is the reception time of the j^{th} successful packet. As we considered $\tau = t'_n$ then $R_i(\tau) = N_i(\tau) = n$. Thus $\Delta_{\tau,i}$ given in (2.7) can be written as

$$\begin{aligned}\Delta_{\tau,i} &= \frac{1}{\tau}\left(Q_1 + \tilde{Q}_{R_i(\tau)} + \sum_{j=2}^{R_i(\tau)} Q_j\right) \\ &= \frac{Q_1 + \tilde{Q}_{R_i(\tau)}}{\tau} + \frac{R_i(\tau) - 1}{\tau}\frac{1}{R_i(\tau) - 1}\sum_{j=2}^{R_i(\tau)} Q_j.\end{aligned}$$

Similarly as in the ATA, this leads to

$$\Delta_i = \lim_{\tau\to\infty} \frac{R_i(\tau) - 1}{\tau}\frac{1}{R_i(\tau) - 1}\sum_{j=2}^{R_i(\tau)} Q_j.$$

Moreover, $R_i(\tau)$ can also be written as $R_i(\tau) = \sup\{j \in \mathbb{N} : \sum_{k=0}^{j-1} Y^k \leq \tau\}$. Thus by using Lemma 2.1 and the fact that $(Y_j)_{j\geq 0}$ is stationary second-order-ergodic, we obtain

$$\lim_{\tau\to\infty} \frac{R_i(\tau) - 1}{\tau} = \frac{1}{\mathbb{E}(Y)}.$$

Using (2.19) and the fact that $(Y_j, T_j)_{j\geq 1}$ is a stationary jointly second-order-ergodic process, we know that

$$\begin{aligned}\lim_{\tau\to\infty}\frac{1}{R_i(\tau)-1}\sum_{j=2}^{R_i(\tau)}Q_j &= \lim_{\tau\to\infty}\frac{1}{R_i(\tau)-1}\sum_{j=2}^{R_i(\tau)}\left(\frac{T_{j-1}^2-T_j^2}{2}+\frac{Y_{j-1}^2}{2}+T_{j-1}Y_{j-1}\right)\\ &= \frac{\mathbb{E}(T^2)-\mathbb{E}(T^2)}{2}+\frac{\mathbb{E}\left(Y^2\right)}{2}+\mathbb{E}\left(TY\right)\\ &= \frac{\mathbb{E}\left(Y^2\right)}{2}+\mathbb{E}\left(TY\right).\end{aligned}$$

Finally, the result above shows that the process $(Q_j)_{j\geq 2}$ is stationary mean-ergodic with a marginal distribution similar to $Q=\frac{Y^2}{2}+TY$. Thus,

$$\lim_{\tau\to\infty}\frac{1}{R_i(\tau)-1}\sum_{j=2}^{R_i(\tau)}Q_j=\mathbb{E}(Q)$$

and

$$\Delta_i=\frac{\mathbb{E}(Q)}{\mathbb{E}(Y)}.$$

This concludes our proof. □

2.3.2 Computing the Average Peak Age

Theorem 2.4. *Assume that the status updating policy* Π *is such that the process* $(Y_j, T_j)_{j\geq 1}$ *is stationary jointly second-order-ergodic with marginal distributions similar to* (Y, T)*. Then the average peak age (PAoI) relative to source* i *is*

$$\Delta_{peak,i}=\mathbb{E}(Y)+\mathbb{E}(T) \tag{2.21}$$

Proof. As in the proof of Theorem 2.3, we define $R_i(\tau)=\sup\{j\in\mathbb{N}; t'_j\leq\tau\}$ the number of successful receptions of source i packets up to time $t=\tau$. From (1.4), we know that

$$\Delta_{peak,i}=\lim_{R_i(\tau)\to\infty}\frac{1}{R_i(\tau)-1}\sum_{j=2}^{R_i(\tau)}K_j.$$

From Fig. 2.2 we observe that $K_j=T_{j-1}+Y_{j-1}$, for all $j\geq 2$. As we are assuming the status updating policy Π to be such that $(Y_j, T_j)_{j\geq 1}$ is stationary jointly second-order-ergodic, then $(K_j)_{j\geq 2}$ is a stationary mean-ergodic process with a marginal distribution similar to $K=Y+T$. Thus,

$$\begin{aligned}\Delta_{peak,i} &= \lim_{R_i(\tau)\to\infty}\frac{1}{R_i(\tau)-1}\sum_{j=1}^{R_i(\tau)}K_j\\ &= \mathbb{E}(K)=\mathbb{E}(Y+T)=\mathbb{E}(Y)+\mathbb{E}(T)\end{aligned}$$

□

It is interesting to observe that, in our computations for the ATA and DTA, we assumed only that the status updating scheme Π considered leads to a stationary second-order-ergodic processes $(\tilde{X}_j, T_j)_{j\geq 1}$ and $(Y_j, T_j)_{j\geq 1}$, respectively. No further assumptions on the status updating scheme or on the probability distributions of the different random variables involved were considered. In the following chapters, we use either ATA or DTA depending on the situation and which approach seems easier to handle.

Part I

Age in the Absence of Noise

3 The gamma Awakening

3.1 Introduction and Main Results

In this chapter[1] , we consider two status-updating policies: Π_1 = { LCFS with preemption and exponentially distributed interarrival times} = { M/G/1/1 with preemption}, and Π_2 = { LCFS-with-preemption-in-waiting with exponentially distributed interarrival times} = { M/G/1/2*}. The main novelty is however the assumption of a gamma distribution for the service time in age of information problems. In this chapter, we use *LCFS with preemption* to refer to Π_1 and *LCFS-with-preemption-in-waiting* to refer to Π_2. As for the reasons behind our choice of the gamma distribution, it is twofold:

- Based on the classic applications of gamma distributions in queuing theory, these distributions can be seen as a reasonable approximation if we want to model relay networks. Indeed, in such network, a transmitter and a receiver are separated by k relays with each relay taking an exponential amount of time to complete transmission to the next hop. This means that the total transmission time is the sum of k independent exponential random variables, which induces a gamma distribution.
- As we will see later, a deterministic random variable can be seen as the limit of a sequence of gamma-distributed random variables. Therefore, we can study the performance of the LCFS-based schemes under deterministic service time by taking the limit of the result obtained for a gamma distributed service time. Although this is an indirect method of calculating (1.2), it is simpler than the direct approach.

As already explained in Section 1.2.3, the two updating schemes studied in this chapter can be described as follows:

[1]The material of this chapter is based on [44, 49].

- **LCFS with preemption**: Any new update will prompt the source to drop the packet being served and start transmitting the newcomer.

- **LCFS-with-preemption-in-waiting**: If the queue is busy, any new update will have to wait in a buffer of size 1. This means that the new update will replace any older packet already waiting to be served.

In both cases, we consider a single source and a single monitor, and we assume the interarrival time process $(X_j)_{j\geq 1}$[2] to be i.i.d. exponentially distributed with rate λ and the service time process $(S_j)_{j\geq 1}$ relative to the transmitted packets to be i.i.d. with a gamma distribution. Moreover, $(X_j)_{j\geq 1}$ and $(S_j)_{j\geq 1}$ are assumed to be independent.

In this chapter, we compute the average age (AoI) and the average peak age (PAoI) for each one of the chosen two status updating policies and compare their performances. We show through simulations that, for an interarrival rate $\lambda \gg \mathbb{E}(S_j)$, the LCFS-with-preemption-in-waiting policy achieves an age lower than the LCFS with preemption. Moreover, we claim that among all gamma distributions, the deterministic service time leads to the worst age performance for the LCFS with preemption scheme. Nonetheless, it leads to the best age performance for the LCFS-with-preemption-in-waiting.

This chapter is organized as follows: In Section 3.2, we present the preliminary results that will be used throughout later. In Section 3.3, we derive the closed-form expressions for both the average age and the average peak age when assuming an LCFS scheme with preemption. In Section 3.4, we compute the formulas for these quantities when considering an LCFS queue without preemption. In these last two sections, the service time is assumed to be gamma distributed. However, in Section 3.5, we calculate the two ages for a deterministic service time for each of the two schemes. Finally, in Section 3.6, we present numerical simulations that validate our theoretical results.

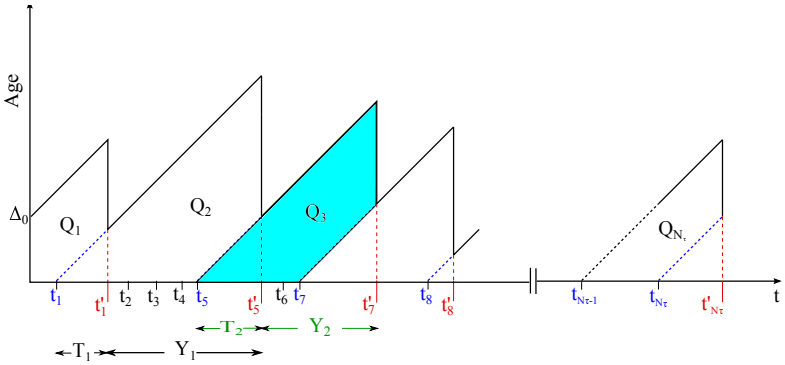
3.2 Preliminaries

3.2.1 General Definitions

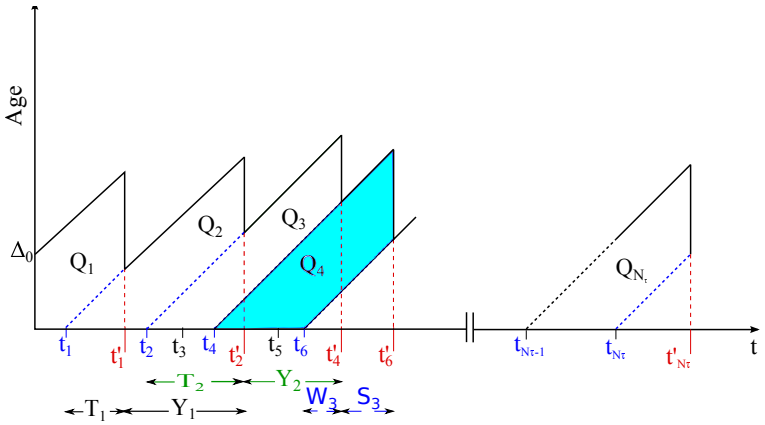
As we have seen in the previous section, our two schemes of interest are LCFS with preemption and LCFS-with-preemption-in-waiting. The variation of the instantaneous age for these two scenarios is given in Figure 3.1. In contrast to Fig. 2.2, these two figures show all the generated packets whether they were successfully delivered or not. For instance, in Fig. 3.1a the updates generated at times t_2, t_3 and t_4 are discarded. In this chapter, we use the notation introduced in Chapter 2. This means

- I_i is the true index of the i^{th} successfully received packet.

[2]The reader should note that it is the interarrival time process and not the *effective* interarrival time process that is assumed to be i.i.d. exponentially distributed.



(a) Age of information for LCFS with preemption scheme



(b) Age of information for LCFS-with-preemption-in-waiting scheme

Figure 3.1 – Variation of the instantaneous age for both schemes

- S_j is the service time of the j^{th} generated packet, with a distribution identical to the random variable S, $f_S(s)$.
- $Y_i = t'_{I_{i+1}} - t'_{I_i}$ is the interdeparture time between the i^{th} and $i+1^{th}$ consecutive successfully received packets.
- $X_j = t_{j+1} - t_j$ is the interarrival time between two consecutive generated packets, distributed similarly as X with $f_X(x) = \lambda e^{-\lambda x}$, $\lambda > 0$.
- $T_i = t'_{I_i} - t_{I_i}$ is the system time of the i^{th} successful packet.

In this chapter, we use the subscript j to indicate random variables related to the j^{th} generated packet (e.g. X_j), and we use the subscript i to denote random variables relative to the i^{th} successful packet (e.g. T_i).

3.2.2 Computing the Average Age

In Chapter 2, we have shown that if the process $(Y_i, T_i)_{i\geq 1}$ is stationary second-order-ergodic, then using the DTA we can write

$$\Delta = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau \Delta(t)dt = \frac{\mathbb{E}(Q)}{\mathbb{E}(Y)} = \lambda_e \mathbb{E}(Q), \tag{3.1}$$

where $\lambda_e = \frac{1}{\mathbb{E}(Y)}$ is the effective update rate, $\mathbb{E}(Y)$ is the expected value of the interdeparture time Y_i at steady state and $\mathbb{E}(Q)$ is the expected value of the area Q_i at steady state. We choose to use the concept of *effective update rate* introduced in [10], instead of computing $\frac{1}{\mathbb{E}(Y)}$, so that the reader is exposed to diverse computation methods related to age of information. Hence, we need to determine these two quantities: λ_e and $\mathbb{E}(Q)$.

Computing the Effective Rate

To calculate the effective rate λ_e, we will use the following lemma:

Lemma 3.1. *Let $\mathbb{1}_{\{j^{th} \text{ generated packet received}\}} = 1$ if the j^{th} generated packet is received and 0 otherwise. If the process $\left(\mathbb{1}_{\{j^{th} \text{ generated packet received}\}}\right)_{j\geq 1}$ is stationary mean-ergodic, then*

$$\lambda_e = \lambda \cdot \mathbb{P}\left(\{\textit{packet is received successfully}\}\right) \tag{3.2}$$

where $\lambda = \frac{1}{\mathbb{E}(X)}$ and $\mathbb{P}\left(\{\textit{packet is received successfully}\}\right)$ is the probability that a packet in the queue will be delivered to the receiver.

Proof. Denoting by $R(\tau)$ the number of successful receptions up to time $t = \tau$, i.e. $R(\tau) = \sup\{i \in \mathbb{N} : \ t'_{I_i} \leq \tau\}$. Let's denote by $M(\tau)$ the number of packets generated in the interval $[0, \tau]$. Then $R(\tau) = \sum_{j=1}^{M(\tau)} \mathbb{1}_{\{j^{th} \text{ generated packet received}\}}$. Thus,

$$\begin{aligned}
\lambda_e = \lim_{\tau\to\infty} \frac{R(\tau)-1}{\tau} &= \lim_{\tau\to\infty} \frac{R(\tau)-1}{R(\tau)} \frac{R(\tau)}{\tau} \\
&= \lim_{\tau\to\infty} \frac{R(\tau)}{\tau} \\
&= \lim_{\tau\to\infty} \frac{M(\tau)}{\tau} \frac{1}{M(\tau)} \sum_{j=1}^{M(\tau)} \mathbb{1}_{\{j^{th} \text{ generated packet received}\}} \\
&\overset{(a)}{=} \frac{\mathbb{E}\left(\{\text{packet is received successfully}\}\right)}{\mathbb{E}(X)} \\
&= \lambda \mathbb{P}\left(\{\text{packet is received successfully}\}\right),
\end{aligned}$$

where equality (a) is obtained by noticing that $M(\tau)$ is a Poisson process with interarrival time distributed as X and using the fact that $\left(\mathbb{1}_{\{j^{th} \text{ generated packet received}\}}\right)_{j\geq 1}$ is stationary mean-ergodic.

□

In the following sections, we will prove that, for the LCFS with preemption and without preemption, the process $\left(\mathbb{1}_{\{j^{th} \text{ generated packet received}\}}\right)_{j\geq 1}$ is stationary mean-ergodic hence (3.1) and Lemma 3.1 can be applied.

Computing $\mathbb{E}(Q_i)$

Using Figures 3.1a and 3.1b, it was shown in Section 2.3 that

$$\mathbb{E}(Q) = \mathbb{E}(Q_i) = \mathbb{E}\left(T_{i-1}Y_{i-1}\right) + \mathbb{E}\left(\frac{Y_{i-1}^2}{2}\right), \tag{3.3}$$

where the choice of i does not matter due to stationarity.

3.2.3 Computing the Average Peak-Age

Another metric of interest is the average peak-age. From Figures 3.1a and 3.1b and Section 2.3, we showed that, for any $i \geq 2$, the average peak-age is given by:

$$\Delta_{peak} = \mathbb{E}(T_{i-1}) + \mathbb{E}(Y_{i-1}) = \mathbb{E}(T) + \mathbb{E}(Y). \tag{3.4}$$

3.2.4 Defining the Service Time

In this chapter, we study two models for the service time: a gamma-distributed service time with parameters (k, θ) and a deterministic service time. Here is a brief description of the gamma distribution.

Definition 3.1. *A random variable S with gamma distribution $\gamma(k, \theta)$ has the following probability density function:*

$$f_S(s) = \frac{s^{k-1}e^{-\frac{s}{\theta}}}{\theta^k \gamma(k)}.$$

The Erlang distribution $E(k, \theta)$ is a special case of the gamma distribution where $k \in \mathbb{N}$.

Such a random variable has a mean of $\mathbb{E}(S) = k\theta$ and a variance $\mathrm{Var}(S) = k\theta^2$. These quantities will come in handy later. Another important property of gamma random variables is given by the following lemma:

Lemma 3.2. *Suppose $S_n \sim \gamma(k_n, \theta_n)$ is a sequence of random variables such that $\mathbb{E}(S_n) = \frac{1}{\mu}$, for some $\mu > 0$. Then the sequence S_n converges in distribution to a deterministic variable Z as k becomes very large, i.e,*

$$S_n \xrightarrow{d} Z, \text{ as } k \to \infty,$$

where $Z = \frac{1}{\mu}$ with probability 1.

The above lemma obviously still holds if $S_n \sim E(k_n, \theta_n)$. This lemma provides additional motivation for studying the average age and the average peak-age under the assumption of a gamma-distributed service time, as we can easily extend the results to the deterministic service time model by letting $k \to \infty$.

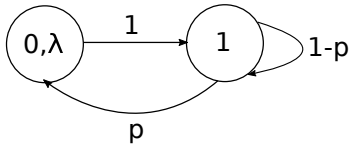


Figure 3.2 – Semi-Markov chain representing the queue for LCFS with preemption

3.3 Age of Information for LCFS with Preemption

In this section, we compute the average age Δ and the average peak-age Δ_{peak} for the Last-Come-First-Served (LCFS) scheme with preemption and a gamma-distributed service time. As we have seen in §1.2.3, in this scenario, if a new packet arrives, any packet being served is preempted, and the new packet is served instead. Hence, the number of packets in the queue can be modeled as a continuous-time two-state semi-Markov chain depicted in Figure 3.2.

The 0-state corresponds to the state where the queue is empty and no packet is being served while the 1-state corresponds to the state where the queue is full and is serving one packet. However, given that the interarrival time between packets is exponentially distributed with rate λ, then we spend an exponential amount of time X in the 0-state before jumping with probability 1 to the other state. Once in the 1-state, two independent clocks are started: The gamma-distributed service time clock of the packet being served and the rate λ memoryless clock of the interarrival time between the current packet and the next one to be generated. We jump back to the 0-state if the service time clock happens to tick before that of the interarrival time. Given that the interarrival times between packets are i.i.d. as well as the service time of each packet, then the probability to jump from the 1-state to the 0-state does not depend on the index of the current packet. Hence, the jump from the 1-state to the 0-state occurs with probability $p = \mathbb{P}(S < X)$, where S is a generic gamma-distributed service time and X is a generic rate λ memoryless interarrival time which is independent of S. Whereas if the interarrival time clock happens to tick before the service time clock, then the current packet being served is preempted and the newly generated packet takes its place in the queue. Therefore, we stay in the 1-state and the two clocks are started anew independently from before. This explains the $1 - p$ probability seen in Figure 3.2 for staying in the 1-state.

Given that the probability p will be useful in the computation of the average age, as well as the average peak-age, we start by deriving its expression here:

$$p = \mathbb{P}(S < X) = \left(\frac{1}{1+\lambda\theta}\right)^k. \tag{3.5}$$

3.3.1 Verifying Convergence

Lemma 3.3. *For the LCFS with preemption status updating scheme, the process* $\left(\mathbb{1}_{\{j^{th} \text{ generated packet received}\}}\right)_{j\geq 1}$ *is i.i.d. Bernoulli(p), where* $p = (1+\lambda\theta)^{-k}$. *By the strong law of large numbers, this process is also mean-ergodic and Lemma 3.1 holds.*

Proof. Let's consider the j^{th} generated packet. The variable $\mathbb{1}_{\{j^{th} \text{ generated packet received}\}}$ depends only on the interarrival time X_j between the j^{th} and the $j+1^{th}$ generated packets and the service time S_j of the j^{th} generated packet. Indeed, the event $\{j^{th}$ generated packet received$\}$ is equivalent to $\{X_j > S_j\}$. Since we assume i.i.d. interarrival times and i.i.d. service times and that, for any j, S_j is independent of $(X_l)_{l\geq 1}$, then the process $\left(\mathbb{1}_{\{j^{th} \text{ generated packet received}\}}\right)_{j\geq 1}$ is i.i.d. Bernoulli(p). □

Lemma 3.4. *Consider a LCFS scheme with preemption. For any $i \geq 1$, the system time, T_i, and interdeparture time, Y_i, relative to the i^{th} successful packet, are independent. Moreover the process $(Y_i)_{i\geq 1}$ is i.i.d. and the process $R(\tau) = \sup\{i \in \mathbb{N} : t'_{I_i} \leq \tau\}$ is a renewal process.*

Proof. The i^{th} successful packet leaves the queue empty hence $Y_i = \hat{X}_i + Z_i$ where $\hat{X}_i = X_i - T_i$ is the remaining of the interarrival time (between the departure of the i^{th} successful packet and the arrival of the next generated one) and Z_i is the time for a new packet to be successfully delivered. Z_i does not overlap with T_i and thus is independent from it. As for $\hat{X}_i$, we also obtain that it is independent of T_i. To prove this, notice that for a successfully received packet i the joint distribution $f_{X_i,T_i}(x,t)$ can be written as

$$f_{X_i,T_i}(x,t) = \begin{cases} 0 & \text{if } x < t \\ \frac{f_{X,S}(x,t)}{\mathbb{P}(S<X)} & \text{if } x > t \end{cases}, \tag{3.6}$$

where X and S are the generic independent interarrival time and service time respectively. Now, using a change of variable we obtain

$$\begin{aligned} f_{\hat{X}_i,T_i}(\hat{x},t) &= f_{X_i-T_i,T_i}(\hat{x},t) = f_{X_i,T_i}(\hat{x}+t,t) \\ &= \begin{cases} 0 & \text{if } \hat{x} < 0 \\ \frac{f_{X,S}(\hat{x}+t,t)}{\mathbb{P}(S<X)} & \text{if } \hat{x} > 0 \end{cases} \\ &= \begin{cases} 0 & \text{if } \hat{x} < 0 \\ \frac{\lambda e^{-\lambda(\hat{x}+t)} f_S(t)}{\mathbb{P}(S<X)} & \text{if } \hat{x} > 0 \end{cases} \\ &= \begin{cases} 0 & \text{if } \hat{x} < 0 \\ \left(\lambda e^{-\lambda \hat{x}}\right) \frac{e^{-\lambda t} f_S(t)}{\mathbb{P}(S<X)} & \text{if } \hat{x} > 0 \end{cases} \\ &= \begin{cases} 0 & \text{if } \hat{x} < 0 \\ h(\hat{x})g(t) & \text{if } \hat{x} > 0 \end{cases}. \end{aligned} \tag{3.7}$$

Moreover, $\hat{X}_i$ is exponential with rate λ since

$$\begin{aligned} \mathbb{P}\left(\hat{X}_i > t\right) &= \mathbb{P}\left(X_i > t + S_i | X_i > S_i\right) \\ &= \frac{\mathbb{P}\left(X_i > t + S_i\right)}{\mathbb{P}(X_i > S_i)} \\ &= \frac{1}{\mathbb{P}(X_i > S_i)} \left(\int_0^\infty e^{-\lambda(t+s)} f_{S_i}(s) \mathrm{d}s \right) \\ &= (1+\lambda\theta)^k \left(\frac{e^{-\lambda t}}{(1+\lambda\theta)^k} \right) \\ &= e^{-\lambda t}. \end{aligned} \tag{3.8}$$

(3.7) and (3.8) show that $\hat{X}_i$ and T_i are indeed independent. Given that $\hat{X}_i$ and Z_i are both independent from T_i, then Y_i and T_i are also independent.

Furthermore, since $Y_{i-1} = \hat{X}_{i-1} + Z_{i-1}$, $\hat{X}_i$ is independent from T_i and the interarrival process is i.i.d. and independent from the i.i.d. service process, then $\hat{X}_i$ and Z_i are independent of Y_{i-1}. This implies that for any $i \geq 1$, Y_{i-1} and Y_i are independent. Moreover, it is clear that the Z_i's have the same distribution (which will be computed later). Since the $\hat{X}_i$'s are exponential with rate λ then the $(Y_i)_{i\geq 1}$ is an i.i.d. process. Given that Y_i is the interval of time between the receptions of two consecutive successful packets, then the number of successfully received packets in the interval $[0, \tau]$, $R(\tau)$, is a renewal process. □

Corollary 3.1. *In the case of a LCFS with preemption scheme and i.i.d. service time process, the average age exists. This implies that the process $(T_i, Y_i)_{i\geq 1}$ is stationary jointly second-moment-ergodic.*

Proof. By Lemma 3.4 we have shown that the process $R(\tau) = \sup\{i \in \mathbb{N} : t'_{I_i} \leq \tau\}$ is a renewal process with $(Y_i)_{i\geq 1}$ being the inter-renewal time process. Thus by defining

$$D_i = \int_{t'_{I_i}}^{t'_{I_{i+1}}} \Delta(t)\mathrm{d}t$$

to be the reward function over the renewal period Y_i, we get using renewal reward theory [16, 58] that

$$\Delta = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau \Delta(t)\mathrm{d}t = \frac{\mathbb{E}(D_i)}{\mathbb{E}(Y_i)} < \infty.$$

This implies that $(T_i, Y_i)_{i\geq 1}$ is stationary jointly second-moment-ergodic. □

3.3.2 Average Age

We start by deriving the expression for the average age. We need to compute two quantities for this purpose: $\mathbb{E}(Q_i)$ and the effective rate λ_e.

Computing the Effective Rate

Using (3.2) and (3.5) we get

$$\lambda_e = \lambda \mathbb{P}\left(\text{packet is received successfully}\right) = \lambda p = \lambda \left(\frac{1}{1+\lambda\theta}\right)^k. \tag{3.9}$$

Computing $\mathbb{E}(Q_i)$

Using (3.3) and Lemma 3.4, we obtain

$$\begin{aligned}\mathbb{E}(Q_i) &= \mathbb{E}\left(T_{i-1}Y_{i-1}\right) + \mathbb{E}\left(\frac{Y_{i-1}^2}{2}\right) \\ &= \mathbb{E}\left(T_{i-1}\right)\mathbb{E}\left(Y_{i-1}\right) + \mathbb{E}\left(\frac{Y_{i-1}^2}{2}\right).\end{aligned} \tag{3.10}$$

Henceforth, we will drop the subscript index because at steady state T_{i-1} and T_i have same the distribution, which is also the case for Y_{i-1} and Y_i. The following lemma will be used to evaluate (3.10):

Lemma 3.5. *Let G be gamma distributed with parameters (k,θ) and F be a rate λ exponential random variable independent of G. Then, conditioned on the event $\{G < F\}$, the distribution of G becomes gamma with parameters $\left(k, \frac{\theta}{1+\lambda\theta}\right)$.*

$$f_{G|G<F}(t) = \frac{t^{k-1}e^{-t\left(\frac{1+\lambda\theta}{\theta}\right)}}{\left(\frac{\theta}{1+\lambda\theta}\right)^k \gamma(k)}. \tag{3.11}$$

Proof. In order to prove this lemma, we will compute the probability density function $f_{G|G<F}$:

$$\begin{aligned}
f_{G|G<F}(t) &= \lim_{\epsilon\to 0} \frac{\mathbb{P}(t \le G < t+\epsilon | G < F)}{\epsilon} \\
&\overset{(a)}{=} \lim_{\epsilon\to 0} \frac{\mathbb{P}(t \le G < t+\epsilon)\mathbb{P}(G < F | t < G < t+\epsilon)}{\epsilon \mathbb{P}(G < F)} \\
&= f_G(t)\frac{\mathbb{P}(F > t)}{\mathbb{P}(G < F)} \\
&\overset{(b)}{=} f_G(t)\frac{e^{-t\lambda}}{p} \\
&= \frac{t^{k-1}e^{-\frac{t}{\theta}}e^{-t\lambda}}{\theta^k\gamma(k)\left(\frac{1}{1+\lambda\theta}\right)^k} \\
&= \frac{t^{k-1}e^{-t\frac{1+\lambda\theta}{\theta}}}{\left(\frac{\theta}{1+\lambda\theta}\right)^k \gamma(k)}
\end{aligned}$$

where (a) is obtained by applying Bayes rule and in (b), p is given by (3.5). □

Proposition 3.1. *At steady state the system time T of a successful packet in a LCFS with preemption scheme is gamma distributed with parameters $\left(k, \frac{\theta}{1+\lambda\theta}\right)$. Therefore,*

$$\mathbb{E}(T) = \frac{k\theta}{1+\lambda\theta}. \tag{3.12}$$

Proof. We first notice that for a given packet i, the event $\{S_i < X_i\}$ is equivalent to the event {packet i was successfully received}. Hence the probability $P = \mathbb{P}(S_i < \alpha | S_i < X_i)$ is the probability that the service time of the i^{th} packet is less than α, given that this packet was successfully transmitted. However, as the service times and interarrival times are i.i.d. , then P does not depend on the index i. Now, as T is the service time of a successful packet then this leads us to

$$\mathbb{P}(T < \alpha) = \mathbb{P}(S_i < \alpha | S_i < X_i) = \mathbb{P}(S < \alpha | S < X), \tag{3.13}$$

where S and X are the generic service and interarrival time respectively. By replacing G by S and F by X in Lemma 3.5, we deduce that the system time T is gamma distributed with parameters $\left(k, \frac{\theta}{1+\lambda\theta}\right)$. □

Now we turn our attention to the distribution of Y, for which we compute its moment generating function. Before going further in our analysis, we state the following lemma.

Lemma 3.6. *Let G be gamma distributed with parameters (k,θ) and F be a rate λ exponential random variable independent of G. If F' is a random variable such that*

$$\mathbb{P}(F' < \alpha) = \mathbb{P}(F < \alpha | F < G),$$

then the moment generating function of F' is given by

$$\phi_{F'}(s) = \mathbb{E}\left(e^{sF'}\right) = \frac{1}{1-p}\left(\frac{\lambda}{\lambda - s} - \frac{\lambda}{\lambda - s}\frac{1}{(1+\theta(\lambda - s))^k}\right), \tag{3.14}$$

where $p = \left(\frac{1}{1+\lambda\theta}\right)^k$.

Proof. We first start by computing the probability density function of F'.

$$\begin{aligned} f_{F'}(t) &= \lim_{\epsilon \to 0} \frac{\mathbb{P}(t \leq F < t+\epsilon | F < G)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\mathbb{P}(t \leq F < t+\epsilon)\mathbb{P}(F < G | t \leq F < t+\epsilon)}{\epsilon \mathbb{P}(F < G)} \\ &= \frac{\lambda e^{-t\lambda}\mathbb{P}(G > t)}{1-p}, \end{aligned}$$

where $p = \left(\frac{1}{1+\lambda\theta}\right)^k$.

So now we can calculate the moment generating function of F'.

$$\begin{aligned} \phi_{F'}(s) &= \int_0^\infty f_{F'}(t) e^{st} \mathrm{d}t \\ &= \int_0^\infty \frac{1}{1-p} \lambda e^{-t\lambda} \mathbb{P}(G > t) e^{st} \mathrm{d}t \\ &= \frac{1}{1-p}\left(\frac{\lambda}{\lambda - s} - \lambda \int_0^\infty \mathbb{P}(G < t) e^{st} \mathrm{d}t\right) \end{aligned}$$

Using integration by parts and the fact that $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(G < t) = f_G(t) = \frac{t^{k-1}e^{-\frac{t}{\theta}}}{\theta^k \gamma(k)}$, we obtain

$$\phi_{F'}(s) = \frac{1}{1-p}\left(\frac{\lambda}{\lambda - s} - \frac{\lambda}{(\lambda - s)(1+\theta(\lambda - s))^k}\right)$$

□

Lemma 3.7. *The moment generating function of Y is given by:*

$$\phi_Y(s) = \frac{\lambda}{\lambda - s\left(1+\theta(\lambda - s)\right)^k}. \tag{3.15}$$

Proof. By observing Figure 3.2, we notice that Y is the smallest time needed to go from the 0-state back to the 0-state. Hence Y can be written as $Y = X + Z$, where X is the generic interarrival time and Z is the time spent in the 1-state before the first jump back to the 0-state. So Z can be written as

$$\begin{aligned} Z &= \begin{cases} S' & \text{with probability } p \\ X'_1 + S' & \text{with probability } (1-p)p \\ X'_1 + X'_2 + S' & \text{with probability } (1-p)^2 p \\ \vdots \end{cases} \\ &= \sum_{j=0}^{M} X'_j + S', \end{aligned} \tag{3.16}$$

where X'_j is such that $\mathbb{P}(X'_j < \alpha) = \mathbb{P}(X < \alpha | X < S)$, S' is such that $\mathbb{P}(S' < \alpha) = \mathbb{P}(S < \alpha | S < X)$ and M is a Geometric(p) random variable that is independent of X'_j and S', and that gives the number of discarded packets before the first successful reception. Applying Lemmas 3.5 and 3.6 on S' and X' respectively and using the fact that M, S' and X'_j are all mutually independent, it follows that

$$\begin{aligned} \phi_Z(s) &= \mathbb{E}\left(e^{s\sum_{j=0}^{M} X'_j}\right) \phi_{S'}(s) \\ &= \mathbb{E}\left(\phi_{X'}(s)^M\right) \left(\frac{1+\lambda\theta}{1+\theta(\lambda-s)}\right)^k \\ &= \sum_{j=0}^{\infty} \phi_{X'}(s)^j p(1-p)^j \left(\frac{1+\lambda\theta}{1+\theta(\lambda-s)}\right)^k \\ &= \frac{\lambda - s}{\lambda - s\left(1+\theta(\lambda-s)\right)^k}. \end{aligned} \tag{3.17}$$

Moreover, since X and W are independent and $\phi_X(s) = \frac{\lambda}{\lambda - s}$, we obtain using (3.17)

$$\phi_Y(s) = \phi_X(s)\phi_Z(s) = \frac{\lambda}{\lambda - s\left(1+\theta(\lambda-s)\right)^k}.$$

□

Now that we have found ϕ_Y, we can compute the first two moments of Y as

$$\mathbb{E}(Y) = \left.\frac{\mathrm{d}\phi_Y(s)}{\mathrm{d}s}\right|_{s=0} = \frac{(1+\lambda\theta)^k}{\lambda}. \tag{3.18}$$

$$\mathbb{E}(Y^2) = \left.\frac{\mathrm{d}^2\phi_Y(s)}{\mathrm{d}s^2}\right|_{s=0} = \frac{2(1+\lambda\theta)^{k-1}}{\lambda^2}\left((1+\lambda\theta)^{k+1} - k\theta\lambda\right). \tag{3.19}$$

Combining these results with (3.12), we obtain,

$$\mathbb{E}(Q_i) = \frac{(1+\lambda\theta)^{2k}}{\lambda^2}. \tag{3.20}$$

Now we are ready to compute the average age.

Theorem 3.1. *The average age in the LCFS with preemption scheme that assumes $\gamma(k,\theta)$ service time is given by:*

$$\Delta = \lambda_e \mathbb{E}(Q_i) = \frac{(1+\lambda\theta)^k}{\lambda}. \tag{3.21}$$

Proof. Using (3.20) and (3.9). □

As we have discussed in Chapter 1, the interarrival rate λ can be a design parameter. If this is the case, then for the LCFS with preemption scheme there exists an optimal rate λ that achieves the minimal average age. This concept is presented in the following lemma.

Lemma 3.8. *Given a LCFS with preemption and gamma(k,θ)-distributed service time, then*

- *if $k > 1$, there exists a finite optimal update rate $\lambda^* < \infty$ that achieves the optimal average age Δ^* with*

$$\lambda^* = \frac{1}{\theta(k-1)} \qquad \Delta^* = k\theta\left(\frac{k}{k-1}\right)^{k-1}.$$

- *if $k \leq 1$, then the average age is a decreasing function of λ and the minimum Δ^* is achieved as $\lambda \to \infty$ with $\Delta^* = 0$.*

Proof. From Theorem 3.1 we know that

$$\Delta = \frac{(1+\lambda\theta)^k}{\lambda}.$$

Thus

$$\frac{\mathrm{d}\Delta}{\mathrm{d}\lambda} = \frac{(\lambda\theta(k-1)-1)(1+\lambda\theta)^{k-1}}{\lambda^2}.$$

Since $\lambda, k, \theta > 0$, then for $k < 1$, $\frac{\mathrm{d}\Delta}{\mathrm{d}\lambda} \leq 0$, $\forall\ \lambda > 0$. This means that Δ is a decreasing function of λ and

$$\Delta^* = \lim_{\lambda\to\infty} \Delta = 0.$$

However, if $k > 1$, the equation $\frac{\mathrm{d}\Delta}{\mathrm{d}\lambda} = 0$ has a unique solution at $\lambda^* = \frac{1}{\theta(k-1)}$. Replacing this value of λ in (3.21) we obtain Δ^*. □

3.3.3 Average Peak-Age

Theorem 3.2. *The average peak-age in the LCFS with preemption scheme that assumes $\gamma(k,\theta)$ service time is given by:*

$$\Delta_{peak} = \mathbb{E}(T) + \mathbb{E}(Y) = \frac{k\theta}{1+\lambda\theta} + \frac{(1+\lambda\theta)^k}{\lambda}. \tag{3.22}$$

Proof. We replace the expectations in (3.4) by their expressions in (3.12) and (3.18). □

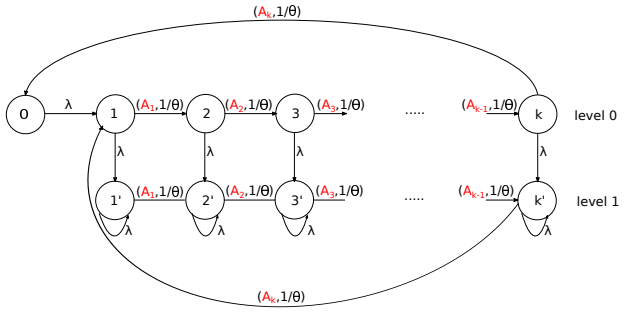


Figure 3.3 – Markov chain representing the queue for LCFS-with-preemption-in-waiting

3.4 Age of Information for LCFS with Preemption in Waiting

Another interesting scheme worth studying is the LCFS-with-preemption-in-waiting. In this scenario, we assume that the queue has a buffer of size 1 and wait for the packet being served to finish before serving a new one. If a new update arrives while serving a packet, it replaces any packet waiting in the buffer. In this section, we derive a closed-form expression for the average age Δ and the average peak-age Δ_{peak} for LCFS-with-preemption-in-waiting, and we assume an Erlang distribution for the service time with parameter (k, θ). An Erlang distribution is simply a special case of the gamma distribution where $k \in \mathbb{N}$. Moreover, an Erlang distribution (k, θ) can be seen as the sum of k independent memoryless random variables A_j, each with rate $\frac{1}{\theta}$. Using this observation, we model the state of the queue as a two-level Markov chain as shown in Figure 3.3.

As in the previous section, we denote the generic rate-λ interarrival time by X and the generic Erlang distributed service time by $S = \sum_{j=1}^{k} A_j$. Using this notation, we notice that the service time can be represented as the succession of k exponential-time steps that need to be accomplished for a successful reception. Hence, a packet in state $j \in \{1, \ldots, k\}$ or $j' \in \{1', \ldots, k'\}$ is a packet completing his j^{th} step out of a total of k. Moreover, the 0-state represents an empty queue, all the states of level 0 represent an empty buffer and those of level 1 represent a full buffer. After spending an exponential amount of time in the 0-state, we can only jump to the 1-state once a new update arrives. Using the memoryless property of the exponential distribution, we can describe the evolution of this packet in the queue as follows: at state $j \in \{1, \ldots, k\}$, two exponential clocks start simultaneously: one clock — denoted A_j— of rate $\frac{1}{\theta}$ and another one — denoted Λ_j— of rate λ. If clock A_j ticks first, then the packet jumps to state $j+1$, and the buffer stays empty. Otherwise it jumps to state j', as now the buffer is full. Whereas, if the packet is at state j' and the A_j clock ticks first, then the packet jumps to state $(j+1)'$ without updating the buffer. However, if the Λ_j ticks first then the packet stays in state j' but we update the buffer with the new arrival.

3.4.1 Verifying Convergence

Lemma 3.9. *For the LCFS-with-preemption-in-waiting status updating scheme with Erlang service time distribution, the Markov chain shown in Fig. 3.3 is ergodic. This means that all processes induced by this chain are ergodic, in particular the process* $\left(\mathbb{1}_{\{j^{th} \text{ generated packet received}\}}\right)_{j\geq 1}$ *is ergodic. Hence, the latter is also mean-ergodic and Lemma 3.1 holds.*

Proof. First we 'uniformize' the Markov chain so that the time spent at each state is exponential with rate $\lambda + \frac{1}{\theta}$, We denote by π_j, $j = 0, 1, \cdots, k$ the steady-state probabilities of the level-0 states in Fig. 3.3, and we denote by π'_l, $l = 1, \cdots, k$ the steady-state probabilities of the level-1 states. Let $q = \frac{1}{1+\lambda\theta}$. The analysis of the embedded chain ([58], chapter 5) gives the system

$$\begin{cases} \pi_1 = \frac{q(1-q)}{q^{k+1}+k(1-q)} \\ \pi_0 = \frac{q^k}{1-q}\pi_1 = \frac{q^{k+1}}{q^{k+1}+k(1-q)} \\ \pi_j = q^{j-1}\pi_1 = \frac{q^j(1-q)}{q^{k+1}+k(1-q)}, \text{ for } j = 2, 3, \cdots k \\ \pi'_l = \frac{1-q^l}{q}\pi_1 = \frac{(1-q^l)(1-q)}{q^{k+1}+k(1-q)}, \text{ for } l = 1, \cdots, k. \end{cases} \tag{3.23}$$

This shows that the embedded Markov chain has a steady-state distribution thus the chain depicted in Fig. 3.3 is ergodic. Moreover we can observe that the event {packet is successfully received} is equivalent to the event {packet passes by the 1-state}, which implies that the process $\left(\mathbb{1}_{\{j^{th} \text{ generated packet received}\}}\right)_{j\geq 1}$ is ergodic. □

Lemma 3.10. *The effective rate of the LCFS-with-preemption-in-waiting status-updating scheme with Erlang*(k, θ) *service-time distribution is*

$$\lambda_e = \frac{\lambda(1+\lambda\theta)^k}{1 + k\lambda\theta(1+\lambda\theta)^k}. \tag{3.24}$$

Proof. We already observed that the event {packet is successfully received} is equivalent to the event {packet passes by the 1-state}. Hence if we 'uniformize' the Markov chain so that the time spent at each state is exponential with rate $\lambda + \frac{1}{\theta}$, we get $\lambda_e = \left(\lambda + \frac{1}{\theta}\right)\pi_1$ where π_1 is the steady-state probability of the 1-state in the 'uniformized' Markov chain. Using (3.23), we get our result. □

3.4.2 Average Age

Theorem 3.3. *The average age in the LCFS-with-preemption-in-waiting scheme assuming Erlang* $E(k, \theta)$ *service time is*

$$\begin{aligned} \Delta = &\frac{k\theta(1+\lambda\theta)^k(2+\lambda\theta+3k\lambda\theta)}{2(1+k\lambda\theta(1+\lambda\theta)^k)} + \frac{2(1-k^2\lambda\theta)}{\lambda(1+k\lambda\theta(1+\lambda\theta)^k)} \\ &+ \frac{k\theta(1+k\lambda\theta+2k)}{1+\lambda\theta+k\lambda\theta(1+\lambda\theta)^{k+1}} \\ &- \frac{1+\lambda\theta+k\lambda\theta}{\lambda(1+\lambda\theta)\left((1+\lambda\theta)^k+k\lambda\theta(1+\lambda\theta)^{2k}\right)}. \end{aligned} \tag{3.25}$$

Proof. Let $q = \frac{1}{1+\lambda\theta}$. As in the previous section, we need to compute the effective rate (given by (3.2)) and $\mathbb{E}(Q_i)$ (given by (3.3)). Since the effective rate is already given by Lemma 3.10, we still need to compute $\mathbb{E}(Q_i)$.

Following the same line of thought as in Section 3.3, we calculate $\mathbb{E}(T_{i-1}Y_{i-1})$ by expressing it as the average of two conditionally independent variables, given some set of events. To this end, we define the family of events $\Psi_j^i = \left\{A_j^i > \Lambda_j^i; \sum_{l=j+1}^k A_l^i < X\right\}$, where $1 \leq j \leq k$. Hence Ψ_j^i is the event that during the service time of the i^{th} successful packet a new update arrived at the j^{th} step of the service time (i.e, state j or j'), then no new update arrived for the remainder of the service time. The superscript (i) is used to indicate that we are dealing with the i^{th} successful packet. For $j = 0$, Ψ_0^i is the event that the i^{th} successful packet leaves the queue empty. Note that for every i, $\{\Psi_j^i, 1 \leq j \leq k\}$ is a partition of the probability space.

It is sufficient to condition on the event Ψ_0^{i-1}, in order to ensure conditional independence between T_{i-1} and Y_{i-1}. This is due to the following fact: given Ψ_0^{i-1}, we know that the $(i-1)^{th}$ successful packet left the queue empty hence we have a situation identical to that of the with preemption case (see Section 3.3) and T_{i-1} and Y_{i-1} are independent. Whereas, given $\overline{\Psi_0^{i-1}}$, the buffer is not empty, hence a new packet will be served directly after the departure of the $(i-1)^{th}$ successful packet. In this case, the interdeparture time Y_{i-1} is simply the service time of the i^{th} successful packet whose value is independent of $T_{i-1} = W_{i-1} + S_{i-1}$, where W_{i-1} the waiting time and S_{i-1} is the service time of the $(i-1)^{th}$ successful packet (see Figure 3.1b).

Although conditioning on Ψ_0^{i-1} is enough to obtain independence between T_{i-1} and Y_{i-1}, we need to condition on the two independent events Ψ_j^{i-1} and Ψ_l^{i-2} in order to be able to calculate the conditional expectation of T_{i-1}. However, it is clear that conditioning on these two events also leads to the independence between T_{i-1} and Y_{i-1}. Hence we obtain

$$\begin{aligned}
&\mathbb{E}(T_{i-1}Y_{i-1}) \\
&= \sum_{j,l=0}^{k} \Big(\mathbb{E}\left(T_{i-1}|\Psi_j^{i-1}\Psi_l^{i-2}\right) \mathbb{E}\left(Y_{i-1}|\Psi_j^{i-1}\Psi_l^{i-2}\right) \\
&\qquad \times \mathbb{P}(\Psi_j^{i-1})\mathbb{P}(\Psi_l^{i-2}) \Big). \qquad (3.26)
\end{aligned}$$

We start by computing $\mathbb{E}\left(T_{i-1}|\Psi_j^{i-1}\Psi_l^{i-2}\right) = \mathbb{E}\left(W_{i-1}|\Psi_j^{i-1}\Psi_l^{i-2}\right) + \mathbb{E}\left(S_{i-1}|\Psi_j^{i-1}\Psi_l^{i-2}\right)$.

The waiting time of the $(i-1)^{th}$ successful packet does not depend on Ψ_j^{i-1}, as they are disjoint in time; but it does depend on Ψ_l^{i-2}. In fact, given Ψ_0^{i-2}, the $(i-1)^{th}$ successful packet will not wait and will start service upon arrival since the $(i-2)^{th}$ successful packet left the queue empty. However, given Ψ_l^{i-2} with $l \neq 0$, the $(i-1)^{th}$ successful packet arrived when the $(i-2)^{th}$ successful packet was at state l or l' of its service time. In order to find the distribution of W_{i-1} conditioned on Ψ_l^{i-2}, we introduce the following event: $\Psi_{l,n}^i = \left\{\sum_{g=1}^n \Lambda_{l,g}^i < A_l^i, \sum_{g=1}^{n+1} \Lambda_{l,g}^i > \sum_{m=l}^k A_m^i\right\}$, where $\{\Lambda_{l,g}^i\}_{g\geq 1}$ is the sequence of interarrival times after the $(i)^{th}$ successful packet

enters state l. Notice that $\Psi_{l,n}^{i}$ is the event that exactly n updates arrived when the i^{th} successful packet was in state l (or l') and then no more updates were generated for the remainder of the service time. Hence $\Psi_l^i = \cup_{n=1}^{\infty}\Psi_{l,n}^{i}$. So conditioned on $\Psi_{l,n}^{(i-2)}$ we have

$$\begin{aligned} W_{i-1} &= \sum_{m=l}^{k} A_m^{(i-2)} - \sum_{g=1}^{n} \Lambda_{l,g}^{(i-2)} \\ &= (A_l^{(i-2)} - \sum_{g=1}^{n} \Lambda_{l,g}^{(i-2)}) + \sum_{m=l+1}^{k} A_m^{(i-2)} \end{aligned} \tag{3.27}$$

It can be shown that, conditioned on $\{\sum_{g=1}^{n}\Lambda_{l,g}^{i} < A_l^i\}$, $(A_l^{(i-2)} - \sum_{g=1}^{n}\Lambda_{l,g}^{(i-2)})$ has an exponential distribution with rate $\frac{1}{\theta}$. This means that under this condition alone, W_{i-1} has the same distribution as the sum of $k-l+1$ independent exponential random variables with rate $\frac{1}{\theta}$. If we further condition on $\left\{\sum_{g=1}^{n+1}\Lambda_{l,g}^{i} > \sum_{m=l}^{k} A_m^i\right\}$ and use Lemma 3.5, we deduce that conditioned on $\Psi_{l,n}^{(i-2)}$, W_{i-1} has a gamma distribution with parameters $\left(k-l+1, \frac{\theta}{1+\lambda\theta}\right)$. Now since $\Psi_l^{i-2} = \cup_{n=1}^{\infty}\Psi_{l,n}^{i-2}$, we conclude that if we condition on Ψ_l^{i-2}, W_{i-1} is distributed as $\gamma\left(k-l+1, \frac{\theta}{1+\lambda\theta}\right)$. Therefore,

$$\mathbb{E}\left(W_{i-1}|\Psi_j^{i-1}\Psi_l^{i-2}\right) = \begin{cases} 0 & \text{if } l = 0 \\ \frac{(k-l+1)\theta}{1+\lambda\theta} & \text{if } l \neq 0 \end{cases}. \tag{3.28}$$

Now we turn our attention to $\mathbb{E}\left(S_{i-1}|\Psi_j^{i-1}\Psi_l^{i-2}\right)$. We first notice that the service time S_{i-1} of the $(i-1)^{th}$ successful packet is independent of its arrival time, given by the event Ψ_l^{i-2}, as we assumed independence between service time and interarrival time. Hence, $\mathbb{E}\left(S_{i-1}|\Psi_j^{i-1}\Psi_l^{i-2}\right) = \mathbb{E}\left(S_{i-1}|\Psi_j^{i-1}\right)$. For the case $j=0$, we obtain

$$\begin{aligned} \mathbb{E}\left(S_{i-1}|\Psi_0^{i-1}\right) &= \mathbb{E}\left(\sum_{m=1}^{k} A_m^{i-1} | \sum_{m=1}^{k} A_m^{i-1} < X\right) \\ &= \frac{k\theta}{1+\lambda\theta} \end{aligned} \tag{3.29}$$

where the last equality is obtained by applying Lemma 3.5 with $G = \sum_{m=1}^{k} A_m^{i-1}$ and

$F = X$. As for the case $j \neq 0$, we get

$$\begin{aligned}
&\mathbb{E}\left(S_{i-1}|\Psi_j^{i-1}\right) \\
&= \mathbb{E}\left(\sum_{m=1}^{k} A_m^{i-1} | A_j^{i-1} > \Lambda_j^{i-1}, \sum_{m=j+1}^{k} A_m^{i-1} < X\right) \\
&= \sum_{m=1}^{j-1} \mathbb{E}(A_m^{i-1}) + \mathbb{E}(A_j^{i-1}|A_j^{i-1} > \Lambda_j^{i-1}) \\
&\quad + \mathbb{E}\left(\sum_{m=j+1}^{k} A_m^{i-1} \middle| \sum_{m=j+1}^{k} A_m^{i-1} < X\right) \\
&\overset{(a)}{=} (j-1)\theta + \frac{\theta(2+\lambda\theta)}{1+\lambda\theta} + \frac{(k-j)\theta}{1+\lambda\theta} \\
&= \frac{\theta(1+k+j\lambda\theta)}{1+\lambda\theta}
\end{aligned} \tag{3.30}$$

where the third term in (a) is obtained by applying Lemma 3.5 with $G = \sum_{m=j+1}^{k} A_m^{i-1}$ and $F = X$. Therefore, combining (3.29) and (3.30) we get,

$$\mathbb{E}\left(T_{i-1}|\Psi_j^{i-1}\Psi_l^{i-2}\right) = \begin{cases} \frac{k\theta}{1+\lambda\theta} & \text{if } l=0, j=0 \\ \frac{\theta(k+1+j\lambda\theta)}{1+\lambda\theta} & \text{if } l=0, j>0 \\ \frac{\theta(2k-l+1)}{1+\lambda\theta} & \text{if } l>0, j=0 \\ \frac{\theta(2k-l+2+j\lambda\theta)}{1+\lambda\theta} & \text{if } l>0, j>0 \end{cases}. \tag{3.31}$$

Now we need to compute $\mathbb{E}\left(Y_{i-1}|\Psi_j^{i-1}\Psi_l^{i-2}\right)$. To this end, observe that Y_{i-1} is independent of Ψ_l^{i-2} given that they do not overlap in time. Moreover, for $j = 0$, the $(i-1)^{th}$ successful packet leaves the queue empty, hence we will need to wait an exponential amount of time X' of rate λ before the i^{th} successful packet arrives and is served directly. Hence, conditioned on Ψ_0^{i-1}, Y_{i-1} has same distribution as $(X' + S)$ with X' and S independent. Whereas for $j \neq 0$, the $(i-1)^{th}$ successful packet leaves the queue with another packet that is waiting in the buffer ready to be served. Thus in this case, Y_{i-1} is simply the service time of the i^{th} successful packet. To sum up,

$$\mathbb{E}\left(Y_{i-1}|\Psi_j^{i-1}\Psi_l^{i-2}\right) = \begin{cases} \frac{1}{\lambda} + k\theta & \text{if } j=0 \\ k\theta & \text{if } j>0 \end{cases} \tag{3.32}$$

To compute $\mathbb{E}(T_{i-1}Y_{i-1})$ we still need the probability $\mathbb{P}(\Psi_j^{i-1})$. For $j > 0$, we use the fact that Ψ_j^{i-1} is the intersection of two independent events and find that $\mathbb{P}(\Psi_j^{i-1}) = \frac{\lambda\theta}{(1+\lambda\theta)^{k-j+1}}$. As for $j = 0$, we have already seen in Section 3.3 that $\mathbb{P}(\Psi_0^{i-1}) = p = \left(\frac{1}{1+\lambda\theta}\right)^k$. These probabilities are independent of the index i hence we can find $\mathbb{P}(\Psi_l^{i-2})$ by replacing j by l in the previous expressions. Combining this results with (3.31), (3.32) we obtain after some tedious calculations

$$\begin{aligned}
\mathbb{E}(T_{i-1}Y_{i-1}) &= \frac{k\theta}{\lambda}(1+k\lambda\theta) + q^k\left(\frac{1-k\lambda\theta(2k+1)}{\lambda^2}\right) \\
&\quad + q^{k+1}\left(\frac{k\theta(1+k\lambda\theta+2k)}{\lambda}\right) - \frac{1}{\lambda^2}q^{2k} - \frac{k\theta}{\lambda}q^{2k+1}
\end{aligned} \tag{3.33}$$

with $q = \frac{1}{1+\lambda\theta}$.
The last term to compute, in order to obtain $\mathbb{E}(Q_i)$, is

$$\mathbb{E}(Y_{i-1}^2) = \mathbb{E}(Y_{i-1}^2|\Psi_0^{i-1})\mathbb{P}(\Psi_0^{i-1}) + \mathbb{E}(Y_{i-1}^2|\overline{\Psi_0^{i-1}})\mathbb{P}(\overline{\Psi_0^{i-1}}).$$

Due to our previous observations, we know that $\mathbb{E}(Y_{i-1}^2|\Psi_0^{i-1}) = \mathbb{E}((X'+S)^2)$ and $\mathbb{E}(Y_{i-1}^2|\overline{\Psi_0^{i-1}}) = \mathbb{E}(S^2)$. Using these facts, we get

$$\mathbb{E}(Y_{i-1}^2) = k\theta^2 + k^2\theta^2 + q^k\left(\frac{2+2k\lambda\theta}{\lambda^2}\right). \tag{3.34}$$

Combining (3.33) and (3.34), we finally get

$$\begin{aligned}\mathbb{E}(Q_i) = &\frac{k\theta(2+\lambda\theta+3k\lambda\theta)}{2\lambda} + 2q^k\left(\frac{1-k^2\lambda\theta}{\lambda^2}\right)\\ &+ q^{k+1}\left(\frac{k\theta(1+k\lambda\theta+2k)}{\lambda}\right) - \frac{1}{\lambda^2}q^{2k} - \frac{k\theta}{\lambda}q^{2k+1}.\end{aligned} \tag{3.35}$$

Finally, replacing $\mathbb{E}(Q_i)$ and λ_e in $\Delta = \lambda_e\mathbb{E}(Q_i)$ by their expressions in (3.35) and (3.24), we obtain our result. □

3.4.3 Average Peak Age

Theorem 3.4. *The average peak-age in the LCFS-with-preemption-in-waiting scheme assuming Erlang $E(k,\theta)$ service time is:*

$$\Delta_{peak} = \frac{1}{\lambda} + 2k\theta - \frac{k\theta}{(1+\lambda\theta)^{k+1}}. \tag{3.36}$$

Proof. Let $q = \frac{1}{1+\lambda\theta}$. We know that $\Delta_{peak} = \mathbb{E}(T_{i-1}) + \mathbb{E}(Y_{i-1})$. We calculate these two terms as follows

$$\begin{aligned}\mathbb{E}(T_{i-1}) &= \sum_{j,l=0}^{k}\mathbb{E}\left(T_{i-1}|\Psi_j^{i-1}\Psi_l^{i-2}\right)\mathbb{P}(\Psi_j^{i-1})\mathbb{P}(\Psi_l^{i-2})\\ &= \frac{1}{\lambda} + k\theta - q^{k+1}\left(\frac{1+\lambda\theta+k\lambda\theta}{\lambda}\right),\end{aligned} \tag{3.37}$$

where we used (3.31) for the last equality. For $\mathbb{E}(Y_{i-1})$ we will only condition on Ψ_0^{i-1}. Thus using (3.32), we get

$$\begin{aligned}\mathbb{E}(Y_{i-1}) &= \mathbb{E}(Y_{i-1}|\Psi_0^{i-1})\mathbb{P}(\Psi_0^{i-1}) + \mathbb{E}(Y_{i-1}|\overline{\Psi_0^{i-1}})\mathbb{P}(\overline{\Psi_0^{i-1}})\\ &= k\theta + \frac{q^k}{\lambda}.\end{aligned} \tag{3.38}$$

Thus, combining the above two results we obtain our result. □

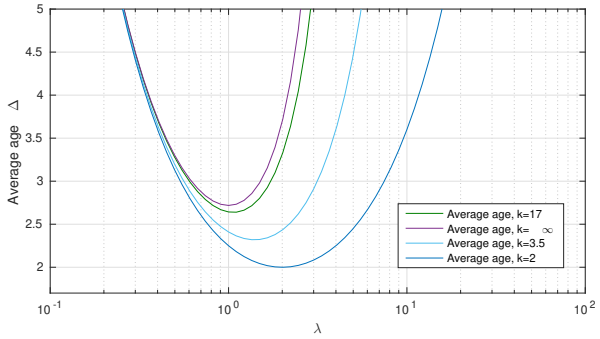


Figure 3.4 – Average age for gamma service time S with $\mathbb{E}(S) = 1$, different k and LCFS with preemption

3.5 Age of Information for Deterministic Service Time

In order to compute the four ages of interest under a deterministic service time assumption, we use Lemma 3.2. For this, we fix the mean of the service times S_n to $\mathbb{E}(S_n) = \frac{1}{\mu}$, for some $\mu > 0$, and let $k \to \infty$. It is beyond the scope of this chapter to show that if $S_n \xrightarrow{d} Z$, as $k \to \infty$, then we also have convergence in the average ages, i.e, $\Delta_{S_n} \to \Delta_Z$. Here Δ_{S_n} refers to the average age corresponding to service time S_n. However, we will use this result to derive the different ages.

3.5.1 LCFS with Preemption

Letting $k \to \infty$ in (3.21) and (3.22), we get

$$\Delta = \frac{e^{\lambda/\mu}}{\lambda} \tag{3.39}$$

$$\Delta_{peak} = \frac{1}{\mu} + \frac{e^{\lambda/\mu}}{\lambda} \tag{3.40}$$

3.5.2 LCFS without Preemption

Letting $k \to \infty$ in (3.25) and (3.36), we get

$$\Delta = \frac{2(2+\rho-\rho^2) - 2e^{-\rho}(1+\rho) + \rho e^{\rho}(2+3\rho)}{2\lambda\,(1+\rho e^{\rho})} \tag{3.41}$$

$$\Delta_{peak} = \frac{1}{\lambda} + \frac{2-e^{-\rho}}{\mu} \tag{3.42}$$

where $\rho = \frac{\lambda}{\mu}$.

3.6 Numerical Results

In this section, we show that the theoretical results obtained in the previous sections match the simulations. We also compare the performance of the two transmission schemes of interest, as well as the effect of the parameter k on each of them. First it

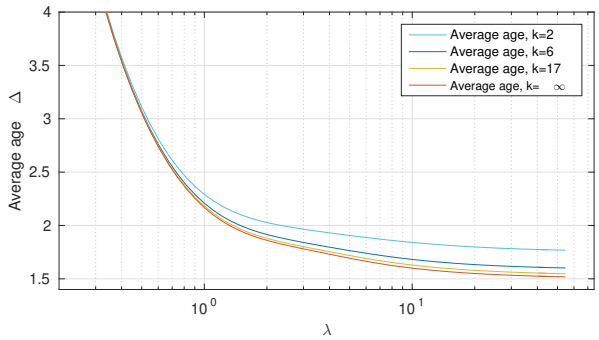


Figure 3.5 – Average age for gamma service time S with $\mathbb{E}(S) = 1$, different k and LCFS-with-preemption-in-waiting

is worth specifying that all simulations were done using gamma distributed service times with all having the same mean $k\theta = 1$, except for the deterministic case where the service time is fixed to 1. Figure 3.4 presents the average age under LCFS with preemption scheme and gamma distributed service time. Two observations can be made based on this plot: (i) the theoretical curves given by (3.21) and (3.39) coincide with the empirical curves, and (ii) as the value of k increases, the average age increases for all values of λ. This means that, under LCFS with preemption, the average age, assuming deterministic service time ($k \to \infty$), is higher than the average age, assuming a regular gamma-distributed service time. In particular, the average age when we assume deterministic service time is higher than the average age assuming memoryless time. This observation can be explained by the fact that the probability of a packet being preempted is given by $1 - p = 1 - \left(\frac{1}{1+\lambda\theta}\right)^k$ (refer to Section 3.3), an increasing function of k. Therefore, as k increases, the receiver will have to wait on average a longer time till a new update is delivered because the preempting rate becomes higher. This analysis is true for any value of λ, hence the phenomenon seen in Figure 3.4.

In a parallel setting, Figure 3.5 presents the average age under LCFS-with-preemption-in-waiting. In this case also, two observations can be made: (i) the theoretical curves given by (3.25) and (3.41) match the empirical results and (ii) as the value of k increases, the average age decreases for almost all λ (except for values close to 0 where all distributions behave similarly). This difference in performance is seen especially at high λ. We explain the intuition behind this behavior. When λ is high ($\lambda \to \infty$), the time where the queue is empty goes to 0 and thus the queue is always transmitting. This also means that on average the waiting time W_{i-1} goes to 0. Given these two observations, we can say that the system time T_{i-1} and the interdeparture time Y_{i-1} will have almost the same distribution as the service time, while being almost independent. Thus

$$\mathbb{E}(Q_i) \overset{\lambda\to\infty}{\longrightarrow} \mathbb{E}(S)^2 + \frac{\mathbb{E}(S^2)}{2}.$$

As for the effective rate λ_e, since the queue is almost always busy, the average rate at which the receiver gets a new update is simply the inverse of the average service

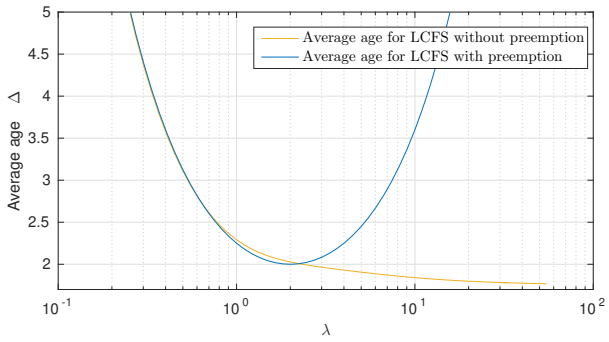


Figure 3.6 – Average age for gamma service time S with $k = 2$ and $\mathbb{E}(S) = 1$

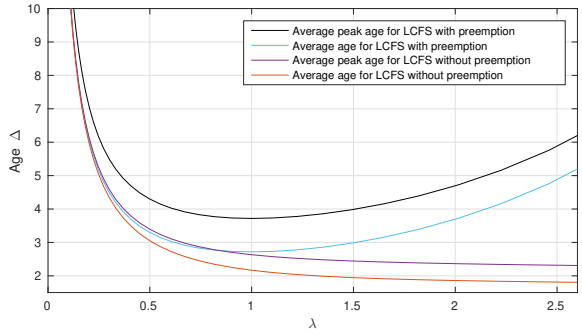


Figure 3.7 – Average age and average peak age for deterministic service time

time, i.e $\lambda_e \overset{\lambda\to\infty}{\longrightarrow} \frac{1}{\mathbb{E}(S)}$. Therefore,

$$\Delta \overset{\lambda\to\infty}{\longrightarrow} \mathbb{E}(S) + \frac{\mathbb{E}(S^2)}{2\mathbb{E}(S)} = \frac{\theta}{2} + \frac{3k\theta}{2}.$$

This result — which is also obtained by taking the limit over λ in (3.25) — is decreasing with k. Hence the behavior seen in Figure 3.5.

Next, we compare the performance of the two transmission schemes in two models: for gamma distributed and deterministic service time. Figure 3.6 shows the average age under LCFS with and without preemption when the service time is taken to be gamma distributed with $k = 2$. In this case we notice that for small λ the two schemes perform similarly. However, for λ's around 1, the LCFS with preemption scheme performs slightly better before being outperformed by the LCFS without preemption scheme at high λ's. In practice, this means that if we use a medium whose service time is modeled as a gamma random variable, the best strategy (among the ones considered) is to not preempt and increase the update generation rate as much as possible. This strategy also applies when the service time is deterministic as seen in Figure 3.7. In fact, we observe that for deterministic service time and for all values of λ, the average age and the average peak-age for the LCFS-with-preemption-in-waiting scheme are smaller than the average age and average peak age for the LCFS with preemption, respectively.

3.7 Conclusion

We have considered the gamma distribution as a model for the service time in status update systems. We have computed and analyzed the average and average peak-age of information under two schemes: LCFS with preemption and LCFS-with-preemption-in-waiting. This have enabled us to evaluate these metrics for the deterministic service time. This suggests that considering gamma distributions for similar problems is a good idea because the gamma distributions (or at least Erlang distributions) are relevant, in practice, as they can be used to model the total service time for relay networks. Moreover, we have shown that, for the LCFS with preemption scheme, there exists an optimal rate λ that achieves the minimal average age. As for the LCFS-with-preemption-in-waiting, simulations indicated that the average age is a decreasing function of λ with an asymptotic value $\Delta^* = \lim_{\lambda\to\infty} \Delta = \frac{\theta}{2}\frac{3k\theta}{2}$.

Another interesting observation is that the average age is highest for the LCFS with preemption scheme when the service time is deterministic, whereas it is lowest for the LCFS-with-preemption-in-waiting among different gamma distributions. Finally, we noticed that depending on the gamma distribution at hand and the value of λ, we could choose either one of the two updating schemes. Whereas for large λ, we should always adopt a LCFS-with-preemption-in-waiting scheme, for small update rates the choice depends on the gamma distribution of the service time.

Status Update in a Multi-stream M/G/1/1 Preemptive Queue

4

4.1 Introduction

In this chapter[1], we assume that an 'observer' (we will call sender), which generates updates according to a Poisson process with rate λ, observes M streams of data. At each generation instant, the source chooses to 'observe' stream i and send its observation (update) of this stream with probability p_i, $i = 1, \ldots, M$. This probability distribution is a design parameter that can be controlled. Moreover, we assume that the system can handle only one update at a time, without any buffer to store incoming updates. This means that whenever a new update is generated and the system is busy, the transmitter preempts the packet being served and starts sending the new update. As we consider a general service-time distribution for the updates, we denote this transmission scheme by M/G/1/1 preemptive queue. It has been shown that for a single-stream source and an exponential update service-time, preemption ensures the lowest average age [5, 35]. However, we have seen in Chapter 3 that under the assumption of gamma-distributed service time, preemption might not be the best policy. We generalize the result of Chapter 3 by deriving in this chapter a closed-form expression for the average age and average peak-age per stream of the multi-stream M/G/1/1 preemptive queue. To this end, we use the detour flow graph method that is also used to find an upper bound on the error probability of a Viterbi decoder (see [56]). As mentioned in Chapter 1, a special case of this problem is studied in [75] where the service time is assumed to be exponentially distributed. In their paper the average age of each stream was obtained in closed form by using a stochastic hybrid system.

Given a fixed total-update rate λ, we also show in this chapter that if we want to decrease the age of a certain stream i with respect to other streams, we need to increase its update rate (by increasing its choice probability p_i) hence decrease the update rates of the other streams. Moreover, if we choose the sum of the ages as our

[1]The material in this chapter is based on [46].

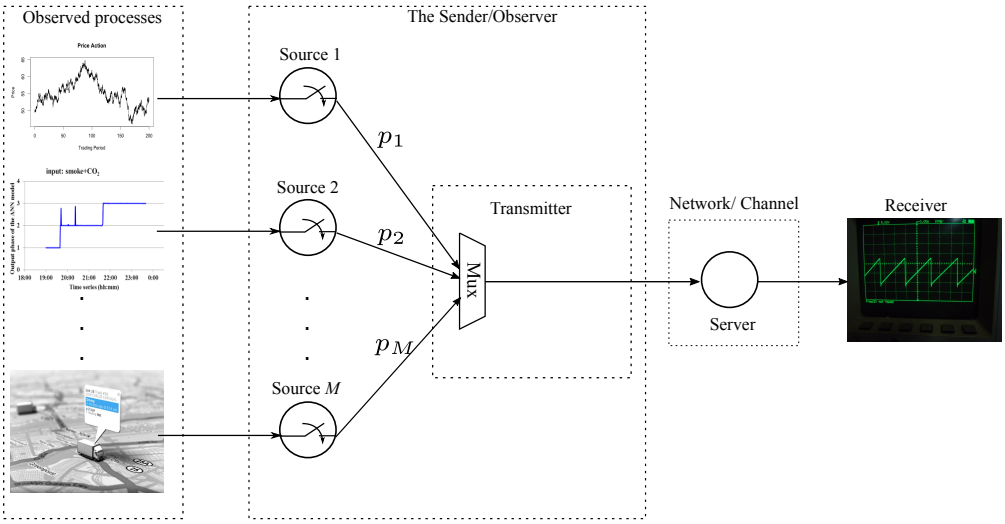


Figure 4.1 – The multi-stream setup: M processes are observed continuously, and the sender generates updates according to a Poisson process with rate λ. At each generation instant, the sender turns on a switch (source) i with probability p_i and sends its observation of the related process through a single server.

performance metric and we seek to minimize it, then we prove that we need to adopt a fair strategy: all streams should be given the same update rate.

This chapter is structured as follows: In Section 4.2, we begin by defining the model and the different variables needed in our study. In Section 4.3, we derive the closed-form expressions of the average age and average peak-age and state the conditions necessary for minimizing the sum of the ages.

4.2 System Model

In this model, a sender generates updates according to a Poisson process with rate λ and sends them through the network. However, we assume that the updates belong to M different streams, each stream i being chosen independently at generation time with probability p_i, $\sum_{i=1}^{M} p_i = 1$. This setup is equivalent to having M independent Poisson sources with rates $\lambda_i = \lambda p_i$, $i = 1, \ldots, M$, and $\lambda = \lambda_1 + \cdots + \lambda_M$ (see [58]). We consider an M/G/1/1 queue with preemption. This means that only one update can be in the system at a time, hence the different streams preempt each other and even the same stream preempts itself. This setup was analyzed in [75], where the authors considered an exponential service time. In this chapter, we assume a service time S with general distribution. Given that the system is symmetric from the point of view of each stream, we focus, without loss of generality, on Stream 1 as the main stream. Thus, unless stated otherwise, all random variables correspond to packets from Stream 1.

In this chapter, we follow the convention where a random variable U with no subscript corresponds to the steady-state version of U_j, which refers to the random variable relative to the j^{th} received packet from Stream 1. To differentiate between streams

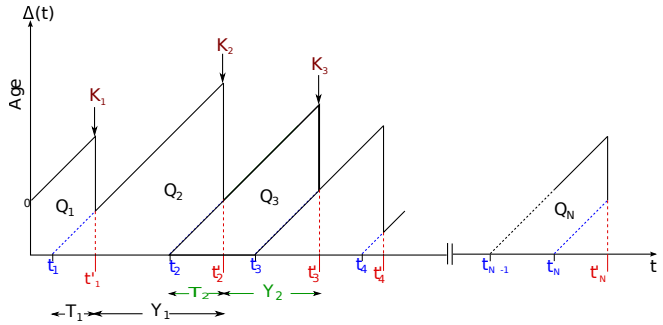


Figure 4.2 – Variation of the instantaneous age of Stream 1 for M/G/1/1 queue with preemption

we use superscripts, thus $U^{(i)}$ corresponds to the steady-state variable U relative to the i^{th} stream.

In this chapter, we also use the notation introduced in Chapter 2. Thus,

- $Y_j = t'_{j+1} - t'_j$ is the interdeparture time between the j^{th} and $j+1^{th}$ successfully received updates from Source 1,
- $X^{(i)}$ is the interarrival time between two consecutive generated updates from stream i, $i = 1, \ldots, M$, (which might or might not be successfully transmitted), so $f_{X^{(i)}}(x) = \lambda_i e^{-\lambda_i x}$,
- S is the service time random variable for any update (from any stream) with distribution $F_S(t)$,
- T_j is the system time, or the time spent by the j^{th} successful update of Source 1 in the queue,
- $R(\tau) = \max\{n : t'_n \leq \tau\}$ is the number of successfully received updates from Stream 1 in the interval $[0, \tau]$.

In our model, we assume that the service time of the updates from the different streams are independent of the interarrival time between consecutive packets (belonging to the same stream or not). These concepts are illustrated in Fig. 4.2, where only successfully transmitted packets from stream 1 are shown.

4.3 Age of a Multi-stream M/G/1/1 Preemptive Queue

We denote by P_λ, the Laplace transform of the service time distribution evaluated at $\lambda = \lambda_1 + \cdots + \lambda_M$, i.e. $P_\lambda = \mathbb{E}\left(e^{-\lambda S}\right)$.

Before stating the main result of this section, we need the following lemmas.

Lemma 4.1. *Let X, Λ and S be three non-negative independent random variables with respective distributions: $f_X(x) = \lambda_1 e^{-\lambda_1 x}$, $f_\Lambda(x) = (\lambda - \lambda_1)e^{-(\lambda-\lambda_1)x}$ and $f_S(t)$, with $\lambda > \lambda_1 > 0$. Let A, Z, B, V be random variables such that $\mathbb{P}(A > t) =$*

$\mathbb{P}(X > t|X < \Lambda)$, $\mathbb{P}(Z > t) = \mathbb{P}(\Lambda > t|X > \Lambda)$, $\mathbb{P}(B > t) = \mathbb{P}(X > t|X < \min(S, \Lambda))$ *and* $\mathbb{P}(V > t) = \mathbb{P}(\Lambda > t|\Lambda < \min(S, X))$. *Then,*

(*i*) $\mathbb{E}\left(e^{sA}\right) = \mathbb{E}\left(e^{sZ}\right) = \frac{\lambda}{\lambda - s}$,

(*ii*) $\mathbb{E}\left(e^{sB}\right) = \mathbb{E}\left(e^{sV}\right) = \frac{\lambda(1-P_{\lambda-s})}{(\lambda-s)(1-P_\lambda)}$,

with P_λ *and* $P_{\lambda-s}$ *being the Laplace transforms of the random variable* S *evaluated at* λ *and* $\lambda - s$*, respectively.*

Proof. We will only prove the result for the variable B, since we can apply the same technique for the others. Denote by $\bar{F}_S(t)$ the complementary CDF of S. Then,

$$\begin{aligned}\mathbb{P}(\min(S, \Lambda) \geq t) &= \mathbb{P}(S \geq t, \Lambda \geq t)\\ &= \mathbb{P}(S \geq t)\,\mathbb{P}(\Lambda \geq t)\\ &= \bar{F}_S(t)e^{-(\lambda-\lambda_1)t}.\end{aligned}$$

Moreover,

$$\begin{aligned}f_B(t) &= \lim_{\epsilon\to 0} \frac{\mathbb{P}(B \in [t, t+\epsilon])}{\epsilon}\\ &= \lim_{\epsilon\to 0} \frac{\mathbb{P}(X \in [t, t+\epsilon]|X \leq \min(S, \Lambda))}{\epsilon}\\ &= \lim_{\epsilon\to 0} \frac{\mathbb{P}(X \in [t, t+\epsilon])\,\mathbb{P}(X \leq \min(S, \Lambda)|X \in [t, t+\epsilon])}{\epsilon\mathbb{P}(X \leq \min(S, \Lambda))}\\ &= \frac{\lambda_1 e^{-\lambda_1 t}\mathbb{P}(\min(S, \Lambda) \geq t)}{\mathbb{P}(X \leq \min(S, \Lambda))}\\ &= \frac{\lambda_1 e^{-\lambda t}\bar{F}_S(t)}{\mathbb{P}(X \leq \min(S, \Lambda))},\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(X \leq \min(S, \Lambda)) &= \int_0^\infty \mathbb{P}(\min(S, \Lambda) \geq t|X = t)\,\lambda_1 e^{-\lambda_1 t}\mathrm{d}t\\ &= \int_0^\infty \lambda_1 e^{-\lambda t}\bar{F}_S(t)\mathrm{d}t = \frac{\lambda_1}{\lambda}(1 - P_\lambda),\end{aligned}$$

where the last equality is obtained using integration by parts. Thus $f_B(t) = \frac{\lambda e^{-\lambda t}\bar{F}_S(t)}{1-P_\lambda}$. Using again integration by parts, we find that

$$\mathbb{E}\left(e^{sB}\right) = \int_0^\infty f_B(t)e^{st}\mathrm{d}t = \frac{\lambda(1 - P_{\lambda-s})}{(\lambda - s)(1 - P_\lambda)}.$$

□

Lemma 4.2. *For the M/G/1/1 queue with preemption described above, the moment generating function of the system time* $T^{(i)}$ *corresponding to a stream* i *is given by*

$$\phi_{T^{(i)}}(s) = \frac{P_{\lambda-s}}{P_\lambda}. \tag{4.1}$$

Note that the right-hand side of (4.1) *does not depend on the chosen stream.*

Proof. Without loss of generality, we will prove Lemma 4.2 for Stream 1. The system time T_j of the j^{th} successfully received packet corresponds to the service time of the j^{th} received packet, given that service was completed before any new arrival (since any new packet from any stream will preempt the current update being served). So, in steady-state, $\mathbb{P}(T > t) = \mathbb{P}\left(S > t | S < \min\left(X^{(1)}, \ldots, X^{(M)}\right)\right)$. Hence, for $L = \min\left(X^{(1)}, \ldots, X^{(M)}\right)$,

$$\begin{aligned} f_T(t) &= \lim_{\epsilon \to 0} \frac{\mathbb{P}(T \in [t, t+\epsilon])}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\mathbb{P}(S \in [t, t+\epsilon] | S < L)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\mathbb{P}(S \in [t, t+\epsilon])\, \mathbb{P}(S < L | S \in [t, t+\epsilon])}{\epsilon \mathbb{P}(S < L)} \\ &= \frac{f_S(t)\mathbb{P}(L > t)}{\mathbb{P}(S < L)} = \frac{f_S(t) e^{-\lambda t}}{\mathbb{P}(S < L)}, \end{aligned}$$

where the last equality is due to the fact that L is exponentially distributed with rate λ. Thus,

$$\phi_T(s) = \mathbb{E}\left(e^{sT}\right) = \int_0^\infty \frac{f_S(t)}{\mathbb{P}(S < L)} e^{-(\lambda - s)t} \mathrm{d}t = \frac{P_{\lambda - s}}{\mathbb{P}(S < L)}.$$

Finally,

$$\begin{aligned} \mathbb{P}(S < L) &= \int_0^\infty f_S(t) \mathbb{P}(L > t)\, \mathrm{d}t = \int_0^\infty f_S(t) e^{-\lambda t} \mathrm{d}t \\ &= P_\lambda. \end{aligned} \tag{4.2}$$

□

Lemma 4.3. *The moment generating function of the interdeparture time of the i^{th} stream, $Y^{(i)}$, is*

$$\phi_{Y^{(i)}}(s) = \frac{\lambda_i P_{\lambda - s}}{\lambda_i P_{\lambda - s} - s}. \tag{4.3}$$

Proof. Without loss of generality, we will prove Lemma 4.3 for Stream 1. We define $L = \min\left(X^{(1)}, \ldots, X^{(M)}\right)$ and $\Lambda = \min\left(X^{(2)}, \ldots, X^{(M)}\right)$. Since L and Λ are the minimum of independent exponential random variables, then they are also exponentially distributed with rates $\lambda = \lambda_1 + \cdots + \lambda_M$ and $\lambda - \lambda_1$, respectively. Fig. 4.3 shows the semi-Markov chain relative to the interdeparture time Y_j between the j^{th} and $j+1^{th}$ received packet of the first stream. When the j^{th} packet leaves the queue, the system enters the idle state q_0 where it waits for a new packet from any stream to be generated. Hence two clocks start: a clock $X^{(1)}$ and a clock Λ. Clock $X^{(1)}$ ticks first with probability $a = \mathbb{P}\left(X^{(1)} < \Lambda\right)$, at which point a new packet from Stream 1 will be generated first and the system goes to state q_1. The value A of the clock when it ticks has distribution $\mathbb{P}(A > t) = \mathbb{P}\left(X^{(1)} > t | X^{(1)} < \Lambda\right)$. Clock Λ ticks first with probability $z = 1 - a = \mathbb{P}\left(\Lambda < X^{(1)}\right)$, at which point a new packet from one of the other $M - 1$ streams is generated first and the system

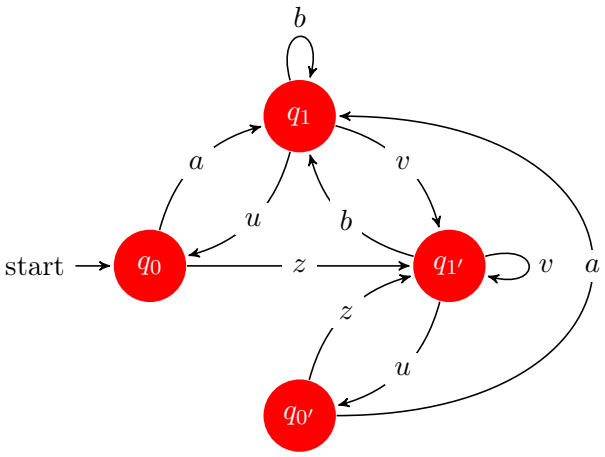


Figure 4.3 – Semi-Markov chain representing the M/G/1/1 interdeparture time for stream 1.

goes to state $q_{1'}$. The value Z of this second clock when it ticks has distribution $\mathbb{P}(Z > t) = \mathbb{P}\left(\Lambda > t | \Lambda < X^{(1)}\right)$.

When the system arrives in state q_1, this means a packet from Stream 1 is beginning its service. Thus, due to the memoryless property of Λ, three clocks start: a service clock S, clock $X^{(1)}$ and clock Λ. The service clock ticks first with probability $u = \mathbb{P}(S < L)$ and its value U has distribution $\mathbb{P}(U > t) = \mathbb{P}(S > t | S < L)$. At this point, the Stream 1 packet currently being served finishes service before any new packet is generated, then the system goes back to state q_0. This ends the interdeparture time Y_j. Whereas, clock $X^{(1)}$ ticks first with probability $b = \mathbb{P}\left(X^{(1)} < \min(S, \Lambda)\right)$ and its value B has distribution $\mathbb{P}(B > t) = \mathbb{P}\left(X^{(1)} > t | X^{(1)} < \min(S, \Lambda)\right)$. At this point, a new stream 1 update is generated before any other update from other streams and preempts the one currently in service. In this case, the system stays in state q_1. The third clock Λ ticks first with probability $v = \mathbb{P}\left(\Lambda < \min\left(S, X^{(1)}\right)\right)$ and its value V has distribution $\mathbb{P}(V > t) = \mathbb{P}\left(\Lambda > t | \Lambda < \min\left(S, X^{(1)}\right)\right)$. At this point, a new update, but not from Stream 1, is generated, preempts the one currently in service and the system switches to state $q_{1'}$.

When the system arrives in state $q_{1'}$, this means a packet not from Stream 1 is beginning its service. Thus, due to the memoryless property of $X^{(1)}$, three clocks start: a service clock S, clock $X^{(1)}$ and clock Λ. As for state q_1, the service clock ticks first with probability u and has value U. At this point, the packet currently being served finishes service before any new packet is generated and the system goes to state $q_{0'}$. Also like before, clock $X^{(1)}$ ticks first with probability b and has value B. At this point, a new Stream 1 update is generated before any other update from other streams and preempts the one currently in service. In this case, the system switches to state q_1. The third clock Λ ticks first with probability v and has value V. At this point, a new update, but not from Stream 1, is generated, preempts the one currently in service and the system stays in state $q_{1'}$.

Finally, when the system arrives in state $q_{0'}$, this means the system is idle but no

update from Stream 1 has been delivered. Given $X^{(1)}$ and Λ are memoryless, the system in state $q_{0'}$ behaves exactly as if it were in state q_0.

From the above analysis, we see that the interdeparture time is given by the sum of the values of the different clocks on the path starting and finishing at q_0. For example, for the path $q_0q_1q_{1'}q_{0'}q_{1'}q_1q_0$ in Fig. 4.3 the interdeparture time

$$Y = A_1 + V_1 + U_1 + Z_1 + B_1 + U_2,$$

where all the random variables in the sum are mutually independent. This value of Y is also valid for the path $q_0q_{1'}q_{0'}q_1q_{1'}q_1q_0$. Hence, Y depends on the variables A_j, B_j, U_j, V_j, Z_j and their number of occurrences and not on the path itself. Therefore, the probability that exactly $(i_1, i_2, i_3, i_4, i_5)$ occurrences of (A, B, U, V, Z) occur, which is equivalent to the probability that

$$Y = \sum_{k=1}^{i_1} A_k + \sum_{k=1}^{i_2} B_k + \sum_{k=1}^{i_3} U_k + \sum_{k=1}^{i_4} V_k + \sum_{k=1}^{i_5} Z_k$$

is given by $a^{i_1}b^{i_2}u^{i_3}v^{i_4}z^{i_5}Q(i_1, i_2, i_3, i_4, i_5)$, where $Q(i_1, i_2, i_3, i_4, i_5)$ is the number of paths with this combination of occurrences. Taking into account the fact that the $\{A_k, B_k, U_k, V_k, Z_k\}$ are mutually independent, the moment generating function of Y is

$$\begin{aligned}
\phi_Y(s) &= \mathbb{E}\left(\mathbb{E}\left(e^{sY} \,|\, (I_1, I_2, I_3, I_4, I_5) = (i_1, i_2, i_3, i_4, i_5)\right)\right) \\
&= \sum_{i_1,i_2,i_3,i_4,i_5} \Big[a^{i_1}b^{i_2}u^{i_3}v^{i_4}z^{i_5}Q(i_1, i_2, i_3, i_4, i_5) \\
&\quad \mathbb{E}\left(e^{s\left(\sum_{k=1}^{i_1} A_k + \sum_{k=1}^{i_2} B_k + \sum_{k=1}^{i_3} U_k + \sum_{k=1}^{i_4} V_k + \sum_{k=1}^{i_5} Z_k\right)}\right)\Big] \\
&= \sum_{i_1,i_2,i_3,i_4,i_5} \Big[a^{i_1}b^{i_2}u^{i_3}v^{i_4}z^{i_5}Q(i_1, i_2, i_3, i_4, i_5) \\
&\quad \mathbb{E}\left(e^{sA}\right)^{i_1} \mathbb{E}\left(e^{sB}\right)^{i_2} \mathbb{E}\left(e^{sU}\right)^{i_3} \mathbb{E}\left(e^{sV}\right)^{i_4} \mathbb{E}\left(e^{sZ}\right)^{i_5}\Big],
\end{aligned} \tag{4.4}$$

where $\{I_1, I_2, I_3, I_4, I_5\}$ are the random variables associated with the number of occurrences of $\{A, B, U, V, Z\}$ respectively.

Moreover, given a directed graph $G = (V, E)$ with algebraic label $L(e)$ on its edges, and a node $u \in V$ with no incoming edges, the transfer function $H(v)$ from u to a node v is the sum over all paths from u to v with each path contributing the product of its edge labels to the sum (see [56, pp. 213–216]). The complete set of transfer functions $\{H(v) : v \in V\}$ can be computed easily by solving the linear equations:

$$\begin{cases} H(u) & = 1 \\ H(w) & = \sum_{w':(w',w)\in E} H(w')L((w', w)), \quad w \neq u. \end{cases}$$

Observe that the sum in (4.4) is nothing but the transfer function from q_0 to $\bar{q}_0$ in the graph shown in Fig. 4.4 with

$$\begin{aligned}
&(D_1, D_2, D_3, D_4, D_5) \\
&= \left(\mathbb{E}\left(e^{sA}\right), \mathbb{E}\left(e^{sB}\right), \mathbb{E}\left(e^{sU}\right), \mathbb{E}\left(e^{sV}\right), \mathbb{E}\left(e^{sZ}\right)\right).
\end{aligned}$$

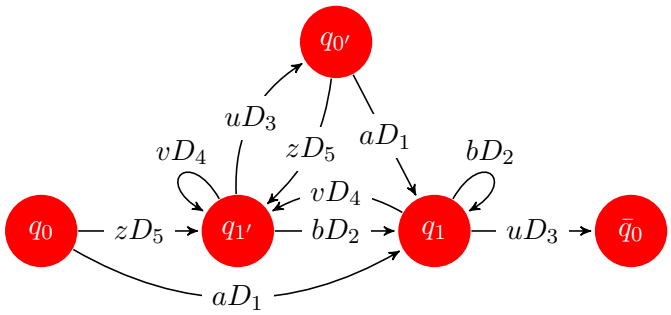


Figure 4.4 – Detour flow graph of the M/G/1/1 interdeparture time for stream 1.

Solving the system of linear equations above yields the transfer function as

$$\begin{aligned} &H(D_1, D_2, D_3, D_4, D_5) \\ &= \sum_{\substack{i_1,i_2,i_3,\\ i_4,i_5}} \left[Q(i_1, i_2, i_3, i_4, i_5) a^{i_1} b^{i_2} u^{i_3} v^{i_4} z^{i_5} D_1^{i_1} D_2^{i_2} D_3^{i_3} D_4^{i_4} D_5^{i_5} \right] \\ &= \frac{uD_3 \left(bD_2 zD_5 + aD_1 - aD_1 vD_4\right)}{\left(1 - bD_2\right)\left(1 - uD_3 zD_5\right) - vD_4 \left(1 + uD_3 aD_1\right)}. \end{aligned} \tag{4.5}$$

Thus

$$\phi_Y(s) = H\left(\mathbb{E}\left(e^{sA}\right), \mathbb{E}\left(e^{sB}\right), \mathbb{E}\left(e^{sU}\right), \mathbb{E}\left(e^{sV}\right), \mathbb{E}\left(e^{sZ}\right)\right).$$

From Lemma 4.1, we know that

$$\mathbb{E}\left(e^{sB}\right) = \mathbb{E}\left(e^{sV}\right) = \frac{\lambda\left(1 - P_{\lambda - s}\right)}{(\lambda - s)\left(1 - P_\lambda\right)} \qquad \text{and} \qquad \mathbb{E}\left(e^{sA}\right) = \mathbb{E}\left(e^{sZ}\right) = \frac{\lambda}{\lambda - s}.$$

Moreover, we notice that U has the same distribution as the system time T hence $\mathbb{E}\left(e^{sU}\right) = \frac{P_{\lambda-s}}{P_\lambda}$. Simple computations show that $a = \frac{\lambda_1}{\lambda}$, $b = \frac{\lambda_1}{\lambda}\left(1 - P_\lambda\right)$, $u = P_\lambda$, $v = \frac{\lambda - \lambda_1}{\lambda}\left(1 - P_\lambda\right)$, $z = \frac{\lambda - \lambda_1}{\lambda}$. Finally, replacing the above expressions into (4.5), we get our result.

□

Theorem 4.1. *Given an M/G/1/1 queue with preemption and service time S and a source generating packets belonging to M streams according to M independent Poisson processes with rates λ_i, $i = 1, \ldots, M$, such that $\lambda = \lambda_1 + \cdots + \lambda_M$, then*

1. *the average age of stream i is given by*

$$\Delta_i = \frac{1}{\lambda_i P_\lambda}, \tag{4.6}$$

2. *and the average peak-age of stream i is given by*

$$\Delta_{peak,i} = \frac{1}{\lambda_i P_\lambda} + \frac{\mathbb{E}\left(Se^{-\lambda S}\right)}{P_\lambda}. \tag{4.7}$$

Proof. Due to the symmetry in the system from a stream point of view, then, without loss of generality, we will prove 4.1 only for Stream 1. The same proof applies for the other $M-1$ streams.

As in Chapter 3, we use the DTA introduced in Section 2.3 to compute the average age. We have shown that the average age for Stream 1 of the M/G/1/1 queue can be also expressed as the sum of the geometric areas Q_i under the instantaneous age curve of Fig. 4.2:

$$\Delta_1 = \lim_{\tau\to\infty} \frac{R(\tau)-1}{\tau} \frac{1}{R(\tau)-1} \sum_{j=2}^{R(\tau)} Q_j = \frac{\mathbb{E}(Q)}{\mathbb{E}(Y)}, \tag{4.8}$$

where Y is the steady-state counterpart of Y_j, Q is the steady-state counterpart of Q_j and the second equality is justified by the fact that $(Y_j, T_j)_{j\geq 1}$ is stationary jointly second-order-ergodic. Using a similar argument as in the proof of Lemma 3.4 and corollary 3.1, and given that the interarrival time of all streams are memoryless, the interdeparture times, Y_j and Y_{j+1}, between two consecutive received updates are i.i.d. Hence $R(\tau)$ forms a renewal process and by [58], $\lim_{\tau\to\infty} \frac{R(\tau)-1}{\tau} = \frac{1}{\mathbb{E}(Y)}$, where Y is the steady-state interdeparture random variable. By defining $D_j = \int_{t'_j}^{t'_{j+1}} \Delta(t)\mathrm{d}t$ to be the reward function over the renewal period Y_j, we obtain using renewal reward theory [16, 58] that

$$\Delta = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau \Delta(t)\mathrm{d}t = \frac{\mathbb{E}(D_j)}{\mathbb{E}(Y_j)} = \frac{\mathbb{E}(Q_j)}{\mathbb{E}(Y_j)} < \infty.$$

This implies that $(T_j, Y_j)_{j\geq 1}$ is stationary jointly second-moment-ergodic.

Moreover, using Fig. 4.2, we see, that by applying the same argument presented in the proof of Lemma 3.4, the variables T_j and Y_j are independent for any $j \geq 1$. Thus, using (2.20),

$$\mathbb{E}\left(Q\right) = \frac{1}{2}\mathbb{E}\left(Y^2\right) + \mathbb{E}\left(TY\right) = \frac{1}{2}\mathbb{E}\left(Y^2\right) + \mathbb{E}\left(T\right)\mathbb{E}\left(Y\right).$$

Therefore,

$$\Delta_1 = \mathbb{E}\left(T\right) + \frac{\mathbb{E}\left(Y^2\right)}{2\mathbb{E}\left(Y\right)} \tag{4.9}$$

From Fig. 4.2 we see that the peak age at the instant before receiving the j^{th} packet is given by

$$K_j = T_{j-1} + Y_{j-1}.$$

Hence, as given by (2.21), at steady state we get

$$\Delta_{peak,1} = \mathbb{E}\left(K\right) = \mathbb{E}\left(T\right) + \mathbb{E}\left(Y\right). \tag{4.10}$$

Using Lemma 4.2, we obtain that

$$\mathbb{E}\left(T\right) = P_\lambda^{-1}\mathbb{E}\left(Se^{-\lambda S}\right).$$

Using Lemma 4.3, we obtain that

$$\mathbb{E}(Y) = (\lambda_1 P_\lambda)^{-1} \quad \text{and} \quad \mathbb{E}\left(Y^2\right) = 2\left(-\frac{\mathbb{E}\left(Se^{-\lambda S}\right)}{\lambda_1 P_\lambda^2} + \frac{1}{\lambda_1^2 P_\lambda^2}\right).$$

Using these expressions in (4.9) and (4.10) we achieve our result for Stream 1. The same argument can be applied to any stream i by replacing λ_1 by λ_i. This proves our theorem. □

Note that, for $M = 1$ and replacing P_λ in (4.6) by the Laplace transform of the gamma distribution evaluated at λ, we recover (3.21). Moreover, if we replace P_λ by the Laplace transform of the exponential distribution evaluated at λ, we recover the expression stated in [75, Theorem 2(a)].

Corollary 4.1. *Let a sender generate updates according to a Poisson process with fixed rate λ. These updates belong to M different streams, each stream i chosen independently with probability p_i at generation time. Then if we use an M/G/1/1 with preemption transmission scheme, we can decrease the average age (and the average peak-age) of a high priority stream k with respect to the other streams by increasing the probability p_k with which it is chosen.*

Proof. From Theorem 4.1, we know that for any two streams i and k, in order to have $\Delta_i < \Delta_k$ or $\Delta_{peak,i} < \Delta_{peak,k}$ we must have $\lambda_i > \lambda_k$. Given that $\lambda_i = \lambda p_i$, $i = 1, \ldots, M$, then we must have $p_i > p_k$. □

Given that the sender generates multiple streams, one interesting performance measure of the system would be the total average age or total average peak age defined respectively as

$$\Delta_{tot} = \sum_{i=1}^{M} \Delta_i, \ \Delta_{peak,tot} = \sum_{i=1}^{M} \Delta_{peak,i}. \tag{4.11}$$

The next theorem gives the distribution over the p_i, $i = 1, \ldots, M$, that minimizes the metrics in (4.11), as well as their minimum achievable value.

Theorem 4.2. *For the M/G/1/1 multi-stream preemptive system described above with fixed total generation rate λ, the optimal strategy that achieves the smallest value for the total average age, Δ_{tot}, and the total average peak-age, $\Delta_{peak,tot}$, is the fair strategy: all streams should have the same generation rate. This means that the probability distribution $\{p_i\}$ over the choices of streams should be the uniform distribution with $p_i = \frac{1}{M}$, $i = 1, \ldots, M$. The optimal values of Δ_{tot} and $\Delta_{peak,tot}$ are given by*

$$\Delta_{tot} = \frac{M^2}{\lambda P_\lambda}, \ \Delta_{peak,tot} = \frac{M^2}{\lambda P_\lambda} + \frac{M\mathbb{E}\left(Se^{-\lambda S}\right)}{P_\lambda} \tag{4.12}$$

Proof. From (4.6), (4.7) and (4.11), we get that

$$\begin{aligned}\Delta_{tot} &= \frac{1}{P_\lambda}\sum_{i=1}^{M}\frac{1}{\lambda_i} = \frac{1}{\lambda P_\lambda}\sum_{i=1}^{M}\frac{1}{p_i}\\ \Delta_{peak,tot} &= \frac{1}{P_\lambda}\sum_{i=1}^{M}\frac{1}{\lambda_i} + \frac{M\mathbb{E}\left(Se^{-\lambda S}\right)}{P_\lambda}\\ &= \frac{1}{\lambda P_\lambda}\sum_{i=1}^{M}\frac{1}{p_i} + \frac{M\mathbb{E}\left(Se^{-\lambda S}\right)}{P_\lambda}\end{aligned} \tag{4.13}$$

Given that λ is fixed, then minimizing Δ_{tot} and $\Delta_{peak,tot}$ over $(p_1, \ldots, p_M)$ is equivalent to minimizing $\sum_{i=1}^{M}\frac{1}{p_i}$. As this is a symmetric convex function, it is minimized when $p_1 = \cdots = p_M = 1/M$ with the value M^2, which proves our theorem. □

From Corollary 4.1 and Theorem 4.2, we see that prioritizing a stream over the others from an age point of view and minimizing the total age are two contradictory objectives.

4.4 Conclusion

In this chapter, we have studied the M/G/1/1 preemptive system with a multi-stream updates sender. The problem solved here is a generalization of the problem solved in Section 3.3 where we considered multiple sources, instead of one, and a general service time distribution, instead of the gamma distribution. We have derived closed-form expressions for the average age and average peak-age using the detour flow graph method. Using these results we have shown that, for a fixed total generation rate, we cannot prioritize one of the streams while minimizing the total age. In fact, we prove that in order to optimize the total age, the source needs to generate all streams at the same rate. This means that no single stream can be given a higher rate, a necessary condition to reduce its age with respect to the other streams.

Content Based Status Updates

5

5.1 Introduction

In the previous chapter, we investigated the case where the sender consists of multiple sources (or streams) that generate updates and send them through one transmitter with one server at its disposal. We assumed that from a content point of view, packets from different streams have the same importance hence no updates belonging to a certain source i are given precedence over packets belonging to other sources. In this chapter[1], we distinguish between the streams and assume that there are sources that generate updates whose content is more important than that of the other sources hence these packets should always be served first.

As in Chapter 3, we assume that updates are generated according to a Poisson process with rate λ, and that the updates belong to two different streams (or sources) where each stream i is chosen independently with probability p_i, $i = 1, 2$. Thus we have two independent Poisson streams with rates $\lambda_1 = \lambda p_1$ and $\lambda_2 = \lambda p_2$. The chapter is divided into two parts:

- In the first part, unlike in Chapter 3, we assume a different transmission policy for each stream. To the best of our knowledge, this model has not been studied before, although it models a natural scenario. In fact, the two independent streams generated by the source can be used to model different types of content carried by the packets of each stream. For example, if the source is a sensor, one stream could carry emergency messages (fire alarm, high pressure, etc.), hence it needs to be always as fresh as possible, whereas the other stream will carry regular updates and hence is not age sensitive. Therefore, it is reasonable to transmit these two streams in a different manner. The regular stream will be transmitted according to a FCFS policy, whereas the high priority stream will be sent by preemption; packets of the high priority stream preempt all packets

[1]The material in this chapter is based on [45, 50].

including packets of their own stream. We further assume that the service time requirements of the two streams are different; a packet of the regular stream is served at rate μ_1, a packet of the priority scheme at rate μ_2.

- In the second part, we assume the same transmission policy for both streams. We use an M/G/1/1 with preemption scheme. However, we consider that a packet from the low-priority (or regular) stream is served according to a service-time distribution similar to that of the random variable S_1, whereas an update from the high-priority stream is served according to a service time distribution identical to that of the variable S_2. We denote by $f_{S_1}(t)$ and $f_{S_2}(t)$ the respective probability density functions (p.d.f) of these service times.

In the first part of this chapter, we will answer the following questions: What should the relation between λ_1, μ_1, λ_2 and μ_2 be for the system to be stable? How does each stream affect the average age of the other one? What are the ages of each stream? To answer these questions, we give a necessary and sufficient condition for the system stability and find the steady-state distribution of the underlying state-space. We also give closed-form expressions for both the average peak-age and a lower bound on the average age of the regular stream, and compare them to the average age of the high-priority stream. For the second part, we will use the detour flow graph method, introduced in Chapter 3, to compute closed-form expressions of the average age and average peak-age relative to the regular and high-priority streams.

This chapter is structured as follows: In Section 5.2, we start by defining the update generation mechanism, common to both models and the different variables needed in our study. In Section 5.3, we study our first model and derive the stability condition of the system and its stationary distribution. The closed-form expressions of the average peak-age and the lower bound on the average age of the regular stream are computed in Section 5.3.2. In Section 5.4, we analyze the second model and compute the average age and average peak-age relative to both streams.

5.2 System Model

We consider a sender that generates packets (or updates) according to a Poisson process of rate λ. Each packet, independently of the previous packets, is of type 1 with probability p_1 and of type 2 with probability $p_2 = 1 - p_1$. We can thus see our sender as consisting of two sources generating two independent Poisson streams $\mathcal{U}_1$ and $\mathcal{U}_2$ with rates $\lambda_1 = \lambda p_1$ and $\lambda_2 = \lambda p_2$ respectively, $\lambda = \lambda_1 + \lambda_2$ (see [58]). As noted in the introduction, the different streams can be used to model packets of different types of content, for example, emergency messages, alerts, error messages, warnings, notices, etc.

We also assume that the updates are sent through a single server (or transmitter) to a monitor. The service times of packets from stream $\mathcal{U}_1$ are i.i.d according to $f_{S_1}(t)$, and those for stream $\mathcal{U}_2$ are i.i.d according to $f_{S_2}(t)$. The difference in service rates between the two streams accounts for the possible difference in compression, packet length, etc., between the two streams. In Section 5.3, the service time of each packet

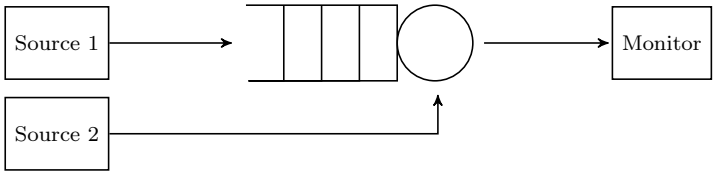


Figure 5.1 – Diagram representing the model with FCFS for the low priority stream.

is considered to be exponentially distributed, with rate μ_1 for stream $\mathcal{U}_1$ and rate μ_2 for stream $\mathcal{U}_2$. However, in Section 5.4 we keep the distributions general.

5.3 FCFS for the Low-Priority Stream

In this model, we constrain the transmitter so that all packets from stream $\mathcal{U}_1$ should be sent. Hence, the server applies a FCFS policy on the packets from stream $\mathcal{U}_1$ with a buffer to save waiting updates. Whereas, we assume that the information carried by stream $\mathcal{U}_2$ is more time sensitive (or has higher priority) hence we aim to minimize its average age. To this end, the transmitter is permitted to perform packet management: In this case, we assume the server applies a preemption policy whenever a packet from $\mathcal{U}_2$ is generated. This means that if a newly generated packet from stream $\mathcal{U}_2$ finds the system busy (serving a packet from $\mathcal{U}_1$ or $\mathcal{U}_2$), the server preempts the update currently in service and starts serving the new packet. On the one hand, if the preempted packet belongs to $\mathcal{U}_1$, this packet is placed back at the head of the $\mathcal{U}_1$-buffer so that it can be served once the system is idle again. On the other hand, if the preempted packet belongs to $\mathcal{U}_2$ then it is discarded. However, if a newly generated $\mathcal{U}_1$-packet finds the system busy serving a $\mathcal{U}_2$-packet, it is placed in the buffer and served when the system becomes idle. This choice of policy for the age sensitive stream is based on the conclusion reached in [6], that for exponentially distributed packet transmission times, the M/M/1/1 with preemption policy is the optimal policy among causal policies. Fig. 5.1 gives a graphical representation of this model.

These ideas are illustrated in part in Fig. 5.2 which also shows the variation of the instantaneous age of stream $\mathcal{U}_1$. In this plot, t_i and D_i refer to the generation and delivery times of the i^{th} packet of stream $\mathcal{U}_1$ while t'_i and D'_i are the start and end times of the i^{th} period during which the system is busy serving packets from stream $\mathcal{U}_2$ only. Notice that for stream $\mathcal{U}_1$ all generated packets are *successful packets*.

5.3.1 System Stability and Stationary Distribution

The fact that we seek to receive all of stream $\mathcal{U}_1$ updates and that stream $\mathcal{U}_2$ has a higher priority and preempts stream $\mathcal{U}_1$ might lead to an unstable system. In order to derive the necessary and sufficient condition for the stability of the system, we study the Markov chain of the number of packets in the system (in service and waiting) shown in Fig. 5.3. In this chain, q_0 is the idle state where the system is completely empty. States q_i, $i > 0$, in the upper row refer to states where the queue is serving a packet from stream $\mathcal{U}_1$, whereas states q'_i, $i > 0$, in the row below correspond to

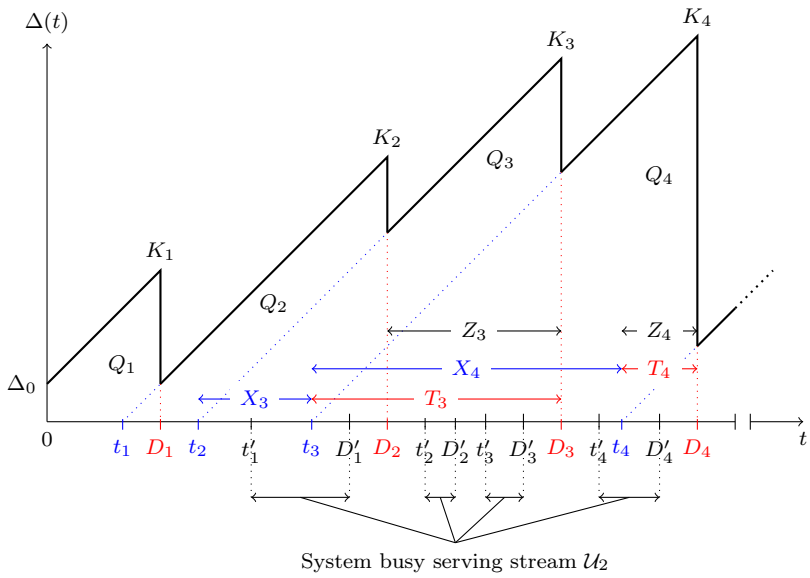


Figure 5.2 – Variation of the instantaneous age of stream $\mathcal{U}_1$.

the queue serving a packet from stream $\mathcal{U}_2$. In both cases, there are $i-1$ stream $\mathcal{U}_1$ updates waiting in the buffer.

The system leaves state q_0 at rate λ_1 to state q_1 when a packet from stream $\mathcal{U}_1$ is generated first and it leaves q_0 at rate λ_2 to state q'_1 when a packet from stream $\mathcal{U}_2$ is generated first. However, when the system enters state q_i, $i > 0$, three exponential clocks start: (i) a clock with rate μ_1, which corresponds to the service time of the stream $\mathcal{U}_1$ packet being served, (ii) a clock with rate λ_1, which corresponds to the generation time of stream $\mathcal{U}_1$ packets and (iii) a clock with rate λ_2, which corresponds to the generation time of stream $\mathcal{U}_2$ packets. If the μ_1-clock ticks first, the system goes to state q_{i-1}: This means that the current stream $\mathcal{U}_1$ packet was delivered and the queue begins the service of the next one in the buffer (if there is any). However, if the λ_1-clock ticks first, a new stream $\mathcal{U}_1$ update is generated and added to the buffer, hence the system goes to state q_{i+1}. Whereas, if the λ_2-clock ticks first, the system preempts the packet currently in service and places it back at the head of the buffer and starts the service of the newly generated stream $\mathcal{U}_2$ update. Thus the system goes to state q'_{i+1}. When the system enters a state q'_i, $i > 0$, two exponential clocks start: the clock with rate λ_1 and a clock with rate μ_2, which corresponds to the service time of a stream $\mathcal{U}_2$ packet. If the λ_1-clock ticks first, the newly generated stream $\mathcal{U}_1$ packet is placed in the buffer and the stream $\mathcal{U}_2$ update is continued to be served. Hence the system goes to state q'_{i+1}. However, if the μ_2-clock ticks first, the stream $\mathcal{U}_2$ packet has finished service and the system starts serving the first stream $\mathcal{U}_1$ packet in the buffer (if there is any). Hence the system goes to state q_{i-1}.

This next theorem gives the necessary and sufficient condition for the above system to be stable, as well as its stationary distribution.

Theorem 5.1. *The system described in Section 5.3 is stable, i.e. the average number*

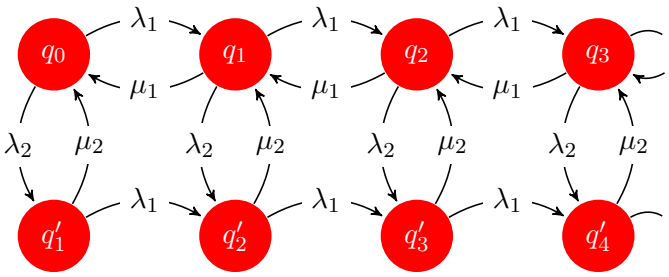


Figure 5.3 – Markov chain governing the number of packets in the system.

of packets in the queue is finite, if and only if

$$\mu_1 > \lambda_1 \left(1 + \frac{\lambda_2}{\mu_2}\right). \tag{5.1}$$

In this case the Markov chain shown in Fig. 5.3 has a stationary distribution $\Pi = [\pi_0, \pi_1, \ldots, \pi_i, \ldots, \pi'_1, \ldots, \pi'_i, \ldots]$, *where* π_i *denotes the stationary probability of state* q_i, $i \geq 0$, *and* π'_i *denotes the stationary probability of state* q'_i, $i > 0$. *This stationary distribution is described by the following system of equations,*

$$\pi_0 = \frac{\mu_2}{\mu_2 + \lambda_2} - \frac{\lambda_1}{\mu_1}, \tag{5.2}$$

$$\begin{bmatrix} \pi_i \\ \pi'_i \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_2 \end{bmatrix} \mathbf{H}^i \begin{bmatrix} \frac{\lambda}{\mu_1} - \frac{\mu_2\lambda_2}{\mu_1(\lambda_1+\mu_2)} \\ \frac{\lambda_2}{\lambda_1+\mu_2} \\ 1 \\ 0 \end{bmatrix} \pi_0, \ i \geq 1 \tag{5.3}$$

where $\lambda = \lambda_1 + \lambda_2$, $\mathbf{H} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{I}_2 & \mathbf{0} \end{bmatrix}$,

$$\mathbf{C} = \begin{bmatrix} 1 + \frac{\lambda}{\mu_1} - \frac{\mu_2\lambda_2}{\mu_1(\mu_2+\lambda_1)} & -\frac{\mu_2\lambda_1}{\mu_1(\mu_2+\lambda_1)} \\ \frac{\lambda_2}{\mu_2+\lambda_1} & \frac{\lambda_1}{\mu_2+\lambda_1} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} -\frac{\lambda_1}{\mu_1} & 0 \\ 0 & 0 \end{bmatrix}.$$

$\mathbf{I}_2$ *is the* 2×2 *identity matrix and* $\mathbf{0}$ *is the* 2×2 *zero matrix.*

Corollary 5.1. *If we define* $N(t)$ *to be the number of stream* $\mathcal{U}_1$ *packets in the system at time* t, *then its moment generating function is* $\phi_{N(t)}$

$$\phi_{N(t)}(s) = \pi_0 \left(\frac{\mu_1 \left(\lambda_1 + \lambda_2 + \mu_2 - \lambda_1 e^s\right)}{\mu_1\mu_2 + \mu_1\lambda_1 - e^s \left(\lambda_1^2 + \lambda_1\lambda_2 + \lambda_1\mu_1 + \lambda_1\mu_2\right) + \lambda_1^2 e^{2s}} \right), \tag{5.4}$$

where π_0 *is given by* (5.2). *Particularly, the expected value of* $N(t)$ *is*

$$\mathbb{E}\left(N(t)\right) = \frac{\lambda_1 \left(2\lambda_2\mu_2 + \lambda_2\mu_1 + \lambda_2^2 + \mu_2^2\right)}{\left(\mu_2 + \lambda_2\right)\left(\mu_1\mu_2 - \lambda_1\left(\mu_2 + \lambda_2\right)\right)}. \tag{5.5}$$

Proof. The distribution given by (5.2) and (5.3) satisfy the detailed balance equations of the Markov chain shown in Fig. 5.3. Moreover, (5.1) is the condition needed to have $\pi_0 > 0$. As for the expression for $\phi_{N(t)}(s)$, it is a consequence of (5.2) and (5.3). The appendix in Section 5.7 presents a full technical version of the proof for Theorem 5.1 and Corollary 5.1. □

The condition in (5.1) can be interpreted as follows: Define the map f from the state-space of the chain as $f(s) = 0$ if s is in $\{q_0, q_1, \ldots\}$ and $f(s) = 1$ if $s \in \{q_1', q_2', \ldots\}$. For each s and s' for which $f(s) = 0$ and $f(s') = 1$ the transition rate from s to s' is the same (λ_2) and similarly the transition rate is μ_2 for s and s' with $f(s) = 1$, $f(s') = 0$. Consequently $F(t) = f(s(t))$, with $s(t)$ being the state at time t, is Markov (which would not be the case for an arbitrary F), and it is easily seen that $F(t) = 0$ a fraction $\phi_0 = \mu_2/(\lambda_2 + \mu_2)$ amount of time, $F(t) = 1$ a fraction $\phi_1 = \lambda_2/(\lambda_2 + \mu_2)$ amount of time. Thus, while the Markov chain in Fig. 5.3 moves right at rate λ_1, it moves left at a rate $\mu_1\phi_0$. The system is stable only if the rate of moving left is larger than the rate of moving right; which gives the condition (5.1).

5.3.2 Ages of Streams $\mathcal{U}_1$ and $\mathcal{U}_2$

Preliminaries

In this section, unless stated otherwise, all random variables correspond to stream $\mathcal{U}_1$. We also follow the convention where a random variable U with no subscript corresponds to the steady-state version of U_j that refers to the random variable relative to the j^{th} received packet from stream $\mathcal{U}_1$. To differentiate between streams, we will use superscripts, which means that $U^{(i)}$ corresponds to the steady-state variable U relative to stream $\mathcal{U}_i$, $i = 1, 2$.

In addition to this, we adopt part of the notation introduced in Chapter 2[2], i.e.,

- $X^{(i)}$ is the interarrival time between two consecutive generated updates from stream $\mathcal{U}_i$, so $f_{X^{(i)}}(x) = \lambda_i e^{-\lambda_i x}$, $i = 1, 2$
- $S^{(i)}$ is the service time random variable of stream $\mathcal{U}_i$ updates, so $f_{S^{(i)}}(t) = \mu_i e^{-\mu_i t}$, $i = 1, 2$,
- T_j is the system time, or the time spent by the j^{th} stream $\mathcal{U}_1$ update in the queue.

In our model, we assume the service time of the updates from the different streams to be independent of the interarrival time between consecutive packets (belonging to the same stream or not).

Deviating from the previous two chapters, to compute the average peak age and the lower bound on the average age relative to stream $\mathcal{U}_1$, in this section we use the interarrival time approach (ATA) presented in Section 2.2. As all generated packets are received, this means that the *effective* interarrival time process $(\tilde{X}_k)_{k\geq 1}$ and the interarrival time process $(X_k)_{k\geq 1}$ are actually the same for stream $\mathcal{U}_1$. Thus, for this low-priority stream, all the formulas derived in Section 2.2 apply by just replacing $\tilde{X}$ by $X^{(1)}$.

[2]In this section, the random variable Y_j is not used to refer to the interdeparture time between the j^{th} and $j + 1^{th}$ received updates. It is used to refer to a special case of service time.

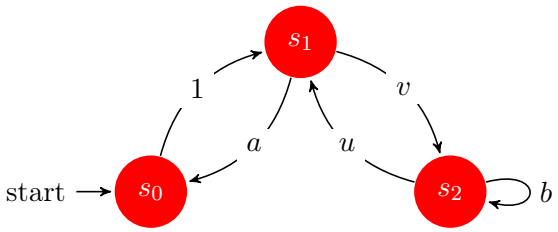


Figure 5.4 – Semi-Markov chain representing the "virtual" service time Y_j.

Analysis of the System

Given the aforementioned description of the model, we can define for each $\mathcal{U}_1$ packet j a "virtual" service time Z_j that could be different from its "physical" service time $S_j^{(1)}$. We define the "virtual" service time Z_j as follows:

$$Z_j = D_j - \max(D_{j-1}, t_j), \tag{5.6}$$

where D_j is the delivery time of the j^{th} packet and t_j is its generation time. Fig. 5.2 shows the "virtual" service time for packets 3 and 4.

For stream $\mathcal{U}_1$, given that the average age calculations seem to be intractable, we compute its average peak age and give a lower bound on its average age. To this end, we first study the steady state "virtual" service time Z.

We define the event

$$\Psi_j = \left\{\text{packet } j \text{ finds the system in state } q_1'\right\}$$

and its complement $\overline{\Psi}_j$. Then, we need the following lemmas.

Lemma 5.1. *Let Y_j be the "virtual" service time of packet j given that this packet does not find the system in state q_1', i.e. $\mathbb{P}\left(Y_j > t\right) = \mathbb{P}\left(Z_j > t|\overline{\Psi}_j\right)$. Then, in steady state,*

$$\phi_Y(s) = \mathbb{E}\left(e^{sY}\right) = \frac{\mu_1(\mu_2 - s)}{s^2 - s(\mu_2 + \mu_1 + \lambda_2) + \mu_1\mu_2}. \tag{5.7}$$

Similarly, let Y_j' be the "virtual" service time of packet j given that this packet finds the system in state q_1', i.e. $\mathbb{P}\left(Y_j' > t\right) = \mathbb{P}\left(Z_j > t|\Psi_j\right)$. Then, in steady state,

$$\phi_{Y'}(s) = \mathbb{E}\left(e^{sY'}\right) = \frac{\mu_1\mu_2}{s^2 - s(\mu_2 + \mu_1 + \lambda_2) + \mu_1\mu_2}. \tag{5.8}$$

Proof. We start by proving (5.7). For this, we use the detour flow graph method. Fig. 5.4 shows the semi-Markov chain relative to the "virtual" service time Y_j of the j^{th} packet of first stream $\mathcal{U}_1$. When the j^{th} packet reaches the head of the buffer, the system is in the idle state s_0. Hence, with probability 1 it goes immediately to state s_1 where it starts serving the j^{th} packet. Due to the memoryless property of the interarrival time of the second stream $X^{(2)}$, two clocks start: a service clock $S^{(1)}$ and a clock $X^{(2)}$. The service clock ticks first with probability $a = \mathbb{P}\left(S^{(1)} < X^{(2)}\right)$ and its value A has distribution $\mathbb{P}\left(A > t\right) = \mathbb{P}\left(S^{(1)} > t|S^{(1)} < X^{(2)}\right)$. At this point, the

stream $\mathcal{U}_1$ packet, currently being served, finishes service before any packet from the other stream is generated, and the system goes back to state s_0. This ends the "virtual" service time Y_j. Clock $X^{(2)}$ ticks first with probability $v = 1 - a = \mathbb{P}\left(X^{(2)} < S^{(1)}\right)$ and its value V has distribution $\mathbb{P}\left(V > t\right) = \mathbb{P}\left(X^{(2)} > t | X^{(2)} < S^{(1)}\right)$. At this point, a new stream $\mathcal{U}_2$ update is generated and preempts the stream $\mathcal{U}_1$ packet currently in service. In this case, the system goes to state s_2, where the preempted stream $\mathcal{U}_1$ update is placed back at the head of the buffer, and the system starts service of the stream $\mathcal{U}_2$ update.

When the system arrives in state s_2, this means a new stream $\mathcal{U}_2$ packet was just generated and is starting its service. Thus, two clocks start: a service clock $S^{(2)}$ and a clock $X^{(2)}$. The service clock ticks first with probability $u = \mathbb{P}\left(S^{(2)} < X^{(2)}\right)$ and its value U has distribution $\mathbb{P}\left(U > t\right) = \mathbb{P}\left(S^{(2)} > t | S^{(2)} < X^{(2)}\right)$. At this point, the packet currently being served finishes service before any new stream $\mathcal{U}_2$ packet is generated, and the system goes back to state s_1 where the j^{th} packet of stream $\mathcal{U}_1$ starts its service again. However, clock $X^{(2)}$ ticks first with probability $b = 1 - u$, and its value $B =$ has distribution $\mathbb{P}\left(B > t\right) = \mathbb{P}\left(X^{(2)} > t | X^{(2)} < S^{(2)}\right)$. At this point, a new stream $\mathcal{U}_2$ update is generated and preempts the one currently in service. In this case, the system stays in state s_2.

From the above analysis, we see that the "virtual" service time is given by the sum of the values of the different clocks on the path starting and finishing at s_0. For example, for the path $s_0s_1s_2s_1s_2s_2s_1s_0$ in Fig. 5.4, the "virtual" service time $Y = V_1 + U_1 + V_2 + B_1 + U_2 + A_1$, where all the random variables in the sum are mutually independent. This value of Y is also valid for the path $s_0s_1s_2s_2s_1s_2s_1s_0$. Hence, Y depends on the variables A_j, B_j, U_j, V_j and their number of occurrences and not on the path itself. Therefore, the probability that exactly (i_1, i_2, i_3, i_4) occurrences of (A, B, U, V) occur, which is equivalent to the probability that

$$Y = \sum_{k=1}^{i_1} A_k + \sum_{k=1}^{i_2} B_k + \sum_{k=1}^{i_3} U_k + \sum_{k=1}^{i_4} V_k$$

is given by $a^{i_1}b^{i_2}u^{i_3}v^{i_4}Q(i_1, i_2, i_3, i_4)$, where $Q(i_1, i_2, i_3, i_4)$ is the number of paths with this combination of occurrences. Taking into account the fact that the $\{A_k, B_k, U_k, V_k\}$ are mutually independent and denoting by $\{I_1, I_2, I_3, I_4\}$ the random variables associated with the number of occurrences of $\{A, B, U, V\}$ respectively, the moment generating function of Y is,

$$\begin{aligned}
\phi_Y(s) &= \mathbb{E}\left(\mathbb{E}\left(e^{sY} | \left(I_1, I_2, I_3, I_4\right) = (i_1, i_2, i_3, i_4)\right)\right) \\
&= \sum_{i_1,i_2,i_3,i_4} \left[a^{i_1}b^{i_2}u^{i_3}v^{i_4}Q(i_1, i_2, i_3, i_4)\mathbb{E}\left(e^{s\left(\sum_{k=1}^{i_1} A_k + \sum_{k=1}^{i_2} B_k + \sum_{k=1}^{i_3} U_k + \sum_{k=1}^{i_4} V_k\right)}\right)\right] \\
&= \sum_{i_1,i_2,i_3,i_4} \left[a^{i_1}b^{i_2}u^{i_3}v^{i_4}Q(i_1, i_2, i_3, i_4)\mathbb{E}\left(e^{sA}\right)^{i_1}\mathbb{E}\left(e^{sB}\right)^{i_2}\mathbb{E}\left(e^{sU}\right)^{i_3}\mathbb{E}\left(e^{sV}\right)^{i_4}\right].
\end{aligned} \tag{5.9}$$

However (5.9) is simply the generating function $H_1(D_1, D_2, D_3, D_4)$ of the detour flow graph shown in Fig. 5.5a, where D_1, D_2, D_3, D_4 are dummy variables (see [56, pp.

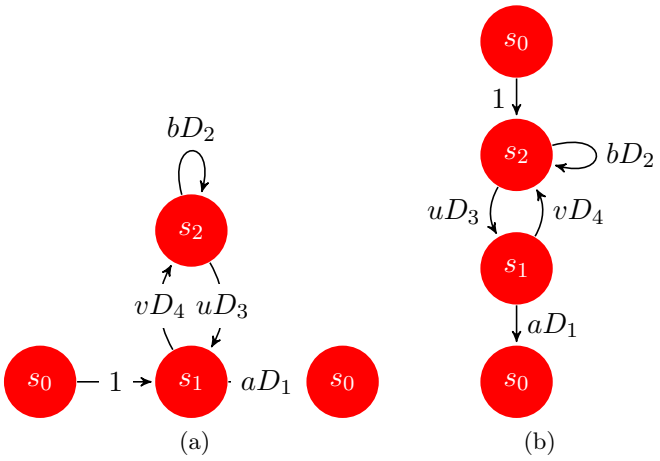


Figure 5.5 – Detour flow graphs for (a) Y and (b) Y'.

213–216]). Simple calculations give

$$\begin{aligned} & H_1(D_1, D_2, D_3, D_4) \\ &= \sum_{i_1,i_2,i_3,i_4} \left[Q(i_1, i_2, i_3, i_4) a^{i_1} b^{i_2} u^{i_3} v^{i_4} D_1^{i_1} D_2^{i_2} D_3^{i_3} D_4^{i_4} \right] \\ &= \frac{aD_1(1 - bD_2)}{1 - bD_2 - uD_3vD_4}. \end{aligned} \tag{5.10}$$

Thus

$$\phi_Y(s) = H_1\left(\mathbb{E}\left(e^{sA}\right), \mathbb{E}\left(e^{sB}\right), \mathbb{E}\left(e^{sU}\right), \mathbb{E}\left(e^{sV}\right)\right).$$

From [75, Appendix A, Lemma 2], we know that A, B, U and V are exponentially distributed with $\mathbb{E}\left(e^{sB}\right) = \mathbb{E}\left(e^{sU}\right) = \frac{\lambda_2+\mu_2}{\lambda_2+\mu_2-s}$ and $\mathbb{E}\left(e^{sA}\right) = \mathbb{E}\left(e^{sV}\right) = \frac{\lambda_2+\mu_1}{\lambda_2+\mu_1-s}$. Simple computations show that $a = \frac{\mu_1}{\mu_1+\lambda_2}$, $b = \frac{\lambda_2}{\mu_2+\lambda_2}$, $u = \frac{\mu_2}{\mu_2+\lambda_2}$, $v = \frac{\lambda_2}{\mu_1+\lambda_2}$. Finally, replacing the above expressions into (5.10), we get our result.

To prove (5.8), we use the same method as before. But in this case, we notice that the j^{th} packet from stream $\mathcal{U}_1$ finds the system busy serving a packet from stream $\mathcal{U}_2$. This translates in the detour flow graph shown in Fig. 5.5b. The generating function of this graph is

$$H_2(D_1, D_2, D_3, D_4) = \frac{aD_1uD_3}{1 - bD_2 - vD_4uD_3}. \tag{5.11}$$

For $(D_1, D_2, D_3, D_4) = \left(\mathbb{E}\left(e^{sA}\right), \mathbb{E}\left(e^{sB}\right), \mathbb{E}\left(e^{sU}\right), \mathbb{E}\left(e^{sV}\right)\right)$ and replacing a, b, u and v by their values in (5.11), we obtain (5.8). □

Lemma 5.2. *The first and second moments of the "virtual" service time Z are given by*

$$\begin{aligned} \mathbb{E}(Z) &= \frac{\lambda_2}{(\lambda_1+\mu_2)(\mu_2+\lambda_2)} + \frac{\lambda_1+\lambda_2+\mu_2}{\mu_1(\lambda_1+\mu_2)}, \\ \mathbb{E}\left(Z^2\right) &= \frac{2\left((\lambda_2+\mu_2)^2(\lambda_2+\mu_2+\lambda_1) + \lambda_2\mu_1(2\lambda_2+\mu_1+2\mu_2)\right)}{\mu_1^2\mu_2(\lambda_1+\mu_2)(\lambda_2+\mu_2)} \end{aligned} \tag{5.12}$$

Proof. For any packet j of stream $\mathcal{U}_1$, conditioning on the event Ψ_j, we get

$$\begin{aligned}\mathbb{E}(Z_j) &= \mathbb{P}(\Psi_j)\,\mathbb{E}(Z_j|\Psi_j) + \mathbb{P}\left(\overline{\Psi}_j\right)\mathbb{E}\left(Z_j|\overline{\Psi}_j\right)\\ &= \mathbb{P}(\Psi_j)\,\mathbb{E}\left(Y'_j\right) + \mathbb{P}\left(\overline{\Psi}_j\right)\mathbb{E}(Y_j),\end{aligned} \tag{5.13}$$

where Y'_j and Y_j are defined as in Lemma 5.1. From Theorem 5.1 we deduce that $\mathbb{P}(\Psi_j) = \pi'_1 = \frac{\lambda_2}{\lambda_1+\mu_2}\pi_0$. So in steady state, $\mathbb{E}(Z) = \pi'_1\mathbb{E}(Y') + (1-\pi'_1)\mathbb{E}(Y)$. Moreover, using (5.7) and (5.8) we get

$$\mathbb{E}(Y) = \frac{\mu_2+\lambda_2}{\mu_1\mu_2},\ \mathbb{E}\left(Y'\right) = \frac{\mu_1+\mu_2+\lambda_2}{\mu_1\mu_2}.$$

Similarly, $\mathbb{E}\left(Z^2\right) = \pi'_1\mathbb{E}\left(Y'^2\right) + (1-\pi'_1)\mathbb{E}\left(Y^2\right)$. Using (5.7) and (5.8) we get

$$\mathbb{E}\left(Y^2\right) = \frac{2\left((\mu_2+\lambda_2)^2+\mu_1\lambda_2\right)}{(\mu_1\mu_2)^2} \text{ and}$$

$$\mathbb{E}\left(Y'^2\right) = \frac{2\left((\mu_1+\mu_2+\lambda_2)^2-\mu_1\mu_2\right)}{(\mu_1\mu_2)^2}.$$

□

Average Peak-Age of Stream $\mathcal{U}_1$

It is worth noting that the system under consideration cannot be seen as an M/G/1 queue with service time distributed as Z, because the "virtual" service times of different packets are correlated. Indeed, if we know that the "virtual" service time of packet j, Z_j, is big, then with very high probability the $(j+1)^{th}$ packet will be generated during the service of the j^{th} packet. Hence, with high probability, Z_{j+1} will be distributed as Y. Whereas, if Z_j is small, then there is a non-negligible probability with which the $(j+1)^{th}$ packet will find the system serving stream $\mathcal{U}_2$. Hence, Z_{j+1} will be distributed as Y'.

Theorem 5.2. *The average peak age of stream $\mathcal{U}_1$ is given by*

$$\Delta_{peak,1} = \frac{1}{\lambda_1} + \frac{2\lambda_2\mu_2+\lambda_2\mu_1+\lambda_2^2+\mu_2^2}{(\mu_2+\lambda_2)\left(\mu_1\mu_2-\lambda_1(\mu_2+\lambda_2)\right)}. \tag{5.14}$$

Proof. As we can deduce from Fig. 5.2 and (2.18), the j^{th} peak $K_j = X_j^{(1)} + T_j$ where $X_j^{(1)}$ is the j^{th} interarrival time for stream $\mathcal{U}_1$ and T_j is the system time of the j^{th} stream $\mathcal{U}_1$ update. At steady state, we get $\Delta_{peak,1} = \mathbb{E}(K) = \mathbb{E}\left(X^{(1)}\right) + \mathbb{E}(T)$. From Little's law we know that $\mathbb{E}(T) = \mathbb{E}(N(t))\,\mathbb{E}\left(X^{(1)}\right)$, with the expected number of stream $\mathcal{U}_1$ packets $\mathbb{E}(N(t))$ given by (5.5) and $\mathbb{E}\left(X^{(1)}\right) = 1/\lambda_1$. □

Lower Bound on the Average Age of Stream $\mathcal{U}_1$

We now compute a lower bound of the average age.

Consider a fictitious system where if a stream $\mathcal{U}_1$ arrival finds the system in state q_1', then the stream $\mathcal{U}_2$ packet that is being served is discarded (and the stream $\mathcal{U}_1$ packet enters service immediately). The instantaneous age process of this fictitious system is pointwise less than the instantaneous age of the true system, consequently its average age lower bounds the true average age. Note that the fictitious system from the point of view of the stream $\mathcal{U}_1$ is M/G/1, with service time distributed like Y in (5.7).

Lemma 5.3. *Assume an M/G/1 queue with interarrival time $X^{(1)}$ exponentially distributed with rate λ_1 and service time Y whose moment generating function is given by* (5.7)*. The service time and the interarrival time are assumed to be independent. Then the distribution of the system time T is*

$$f_T(t) = C_1 e^{-\alpha_1 t}(\mu_2 - \alpha_1) - C_1 e^{-\alpha_2 t}(\mu_2 - \alpha_2),\ t \geq 0, \tag{5.15}$$

where $\alpha_1, \alpha_2 > 0$ are the roots of the quadratic expression

$$s^2 - s(\mu_1 + \mu_2 + \lambda_2 - \lambda_1) + \mu_1\mu_2 - \lambda_1\mu_2 - \lambda_1\lambda_2,$$

$$C_1 = \frac{(1-\rho)\mu_1}{\alpha_2 - \alpha_1},$$

and $\rho = \lambda_1 \mathbb{E}(Y) = \frac{\lambda_1(\mu_2+\lambda_2)}{\mu_1\mu_2}$.

Proof. From [13, p. 166], we know that the Laplace transform of the system time T is

$$\mathbb{E}\left(e^{-sT}\right) = \frac{(1-\rho)s\phi_Y(-s)}{s - \lambda_1(1 - \phi_Y(-s))}.$$

Replacing $\phi_Y(-s)$ by its expression in (5.7) we get

$$\begin{aligned}
\mathbb{E}\left(e^{-sT}\right) &= \frac{(1-\rho)\mu_1(\mu_2 + s)}{s^2 + s(\mu_1 + \mu_2 + \lambda_2 - \lambda_1) + \mu_1\mu_2 - \lambda_1\mu_2 - \lambda_1\lambda_2} \\
&= \frac{(1-\rho)\mu_1(\mu_2 + s)}{(s - s_1)(s - s_2)} \\
&= s\frac{(1-\rho)\mu_1}{(s - s_1)(s - s_2)} + \frac{(1-\rho)\mu_1\mu_2}{(s - s_1)(s - s_2)},
\end{aligned} \tag{5.16}$$

where s_1 and s_2 are two real roots of the quadratic equation

$$s^2 + s(\mu_1 + \mu_2 + \lambda_2 - \lambda_1) + \mu_1\mu_2 - \lambda_1\mu_2 - \lambda_1\lambda_2.$$

Moreover, due to condition (5.1),

$$s_1 + s_2 = -\mu_1 - \mu_2 - \lambda_2 + \lambda_1 < 0$$

and

$$s_1 s_2 = \mu_1\mu_2 - \lambda_1\mu_2 - \lambda_1\lambda_2 > 0.$$

This proves that both roots s_1 and s_2 are negative. Let

$$G(s) = \frac{(1-\rho)\mu_1}{(s-s_1)(s-s_2)},$$

and $g(t)$ its inverse Laplace transform. Using the initial value theorem:

$$g(0^+) = \lim_{s\to\infty} sG(s) = 0. \tag{5.17}$$

Using (5.17) and the expression of $G(s)$, (5.16) can be written as

$$\mathbb{E}\left(e^{-sT}\right) = sG(s) - g(0^+) + \mu_2 G(s). \tag{5.18}$$

Therefore, the probability density function of the system time $f_T(t)$ (which is the inverse Laplace transform of $\mathbb{E}\left(e^{-sT}\right)$) is

$$f_T(t) = g'(t) + \mu_2 g(t). \tag{5.19}$$

By partial fraction expansion,

$$G(s) = \frac{C_1}{s-s_1} - \frac{C_1}{s-s_2},$$

where $C_1 = \frac{(1-\rho)\mu_1}{s_1-s_2}$. Denoting $\alpha_1 = -s_1 > 0$ and $\alpha_2 = -s_2 > 0$, we get

$$G(s) = \frac{C_1}{s+\alpha_1} - \frac{C_1}{s+\alpha_2}, \text{ and } C_1 = \frac{(1-\rho)\mu_1}{\alpha_2-\alpha_1}.$$

Thus,

$$g(t) = C_1 e^{-\alpha_1 t} - C_1 e^{-\alpha_2 t},$$

and

$$f_T(t) = C_1 e^{-\alpha_1 t}(\mu_2 - \alpha_1) - C_1 e^{-\alpha_2 t}(\mu_2 - \alpha_2).$$

□

From [34] and (2.16), we know that the average age of the M/G/1 queue with interarrival time $X^{(1)}$ and service time Y is

$$\Delta_{LB} = \lambda_1 \left(\frac{1}{2}\mathbb{E}\left(X_j^{(1)^2}\right) + \mathbb{E}\left(T_j X_j^{(1)}\right)\right), \tag{5.20}$$

where for the j^{th} packet we have $T_j = (T_{j-1} - X_j^{(1)})^+ + Y_j$, $f(x) = (x)^+ = x\mathbb{1}_{\{x\geq 0\}}$ and $\mathbb{1}_{\{.\}}$ is the indicator function. So $\mathbb{E}\left(T_j X_j^{(1)}\right)$ becomes

$$\mathbb{E}\left(T_j X_j^{(1)}\right) = \mathbb{E}\left(X_j^{(1)}(T_{j-1} - X_j^{(1)})^+\right) + \mathbb{E}\left(Y_j\right)\mathbb{E}\left(X_j^{(1)}\right), \tag{5.21}$$

where the second term is due to the independence between Y_j and $X_j^{(1)}$.

Proposition 5.1.

$$\begin{aligned}
&\mathbb{E}\left(X_j^{(1)}(T_{j-1}-X_j^{(1)})^+\right)\\
&=\frac{\lambda_1\mu_2+2\lambda_1\lambda_2}{\mu_1^2(\mu_1\mu_2-\lambda_1(\mu_2+\lambda_2))}+\frac{\lambda_2\lambda_1}{\mu_2}\left(\frac{(\mu_2+\mu_1+\lambda_2)^2-2\mu_1\mu_2}{\mu_1^2(\mu_2+\lambda_1)(\mu_1\mu_2-\lambda_1(\mu_2+\lambda_2))}\right)\\
&+\frac{\lambda_2\lambda_1}{\mu_2}\left(\frac{2\mu_2\lambda_1(\mu_1+\lambda_2)+\lambda_2(\lambda_1^2+\mu_2)}{\mu_1^2(\mu_2+\lambda_1)^2(\mu_1\mu_2-\lambda_1(\mu_2+\lambda_2))}\right).
\end{aligned} \tag{5.22}$$

Proof. Given that T_{j-1} and $X_j^{(1)}$ are independent then

$$\begin{aligned}
&\mathbb{E}\left(X_j^{(1)}(T_{j-1}-X_j^{(1)})^+\right)\\
&=\int_0^\infty\int_x^\infty x(t-x)f_T(t)\lambda_1 e^{-\lambda_1 x}\mathrm{d}t\mathrm{d}x
\end{aligned}$$

Replacing $f_T(t)$ by its value in (5.15) and using the fact that

$$\begin{aligned}
\alpha_1+\alpha_2&=\mu_1+\mu_2+\lambda_2-\lambda_1,\\
\alpha_1\alpha_2&=\mu_1\mu_2-\lambda_1\mu_2-\lambda_1\lambda_2,
\end{aligned}$$

we get (5.22) after some computations. □

Theorem 5.3.

$$\begin{aligned}
\Delta_{LB}&=\frac{1}{\lambda_1}+\frac{\mu_2+\lambda_2}{\mu_1\mu_2}+\frac{\lambda_1^2\mu_2+2\lambda_1^2\lambda_2}{\mu_1^2(\mu_1\mu_2-\lambda_1(\mu_2+\lambda_2))}\\
&+\frac{\lambda_2\lambda_1^2}{\mu_2}\left(\frac{(\mu_2+\mu_1+\lambda_2)^2-2\mu_1\mu_2}{\mu_1^2(\mu_2+\lambda_1)(\mu_1\mu_2-\lambda_1(\mu_2+\lambda_2))}\right)\\
&+\frac{\lambda_2\lambda_1^2}{\mu_2}\left(\frac{2\mu_2\lambda_1(\mu_1+\lambda_2)+\lambda_2(\lambda_1^2+\mu_2)}{\mu_1^2(\mu_2+\lambda_1)^2(\mu_1\mu_2-\lambda_1(\mu_2+\lambda_2))}\right).
\end{aligned} \tag{5.23}$$

This is also a lower bound on the true average age of stream $\mathcal{U}_1$ packets.

Proof. Using (5.22),

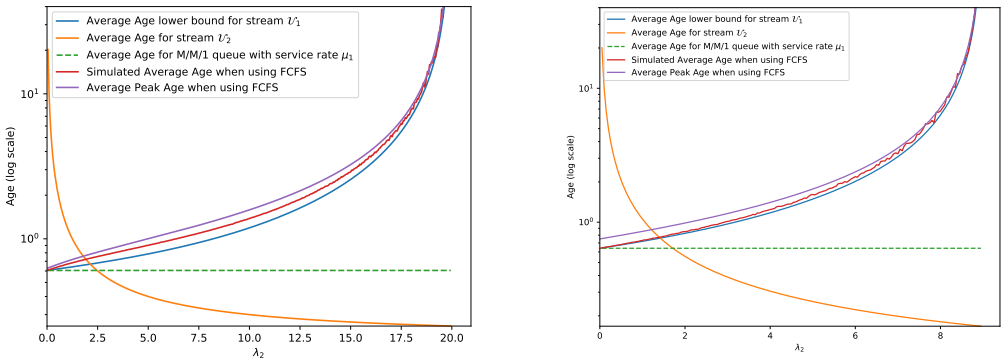
$$\mathbb{E}(Y_j)=\mathbb{E}(Y)=\frac{\mu_2+\lambda_2}{\mu_1\mu_2}\quad\text{and}\quad\mathbb{E}\left(X_j^{(1)}\right)=\mathbb{E}\left(X^{(1)}\right)=\frac{1}{\lambda_1},$$

we can find a closed-form expression for $\mathbb{E}\left(T_jX_j^{(1)}\right)$. Replacing this expression in (5.20) and using the fact that $\mathbb{E}\left(X_j^{(1)^2}\right)=\frac{2}{\lambda_1^2}$, we obtain a closed-form expression of the average age Δ_{LB} of an M/G/1 queue with interarrival time $X^{(1)}$ and service time Y. □

Average Age of Stream $\mathcal{U}_2$

By design, stream $\mathcal{U}_2$ is not interfered at all by stream $\mathcal{U}_1$ hence behaves like a traditional M/M/1/1 with preemption queue with generation rate λ_2 and service rate μ_2. The average age of this stream was computed in [35] to be

$$\Delta_{\mathcal{U}_2}=\frac{1}{\mu_2}+\frac{1}{\lambda_2}. \tag{5.24}$$



(a) $\mu_1 = 10$, $\mu_2 = 5$, $\lambda_1 = 2$ and $\lambda_2 < \frac{\mu_1\mu_2}{\lambda_1} - \mu_2 = 20$. (b) $\mu_1 = 6$, $\mu_2 = 18$, $\lambda_1 = 4$ and $\lambda_2 < 9$.

Figure 5.6 – Plot of the average age for stream $\mathcal{U}_2$ and average peak age and lower bound on the average age for stream $\mathcal{U}_1$.

5.3.3 Numerical Results

Fig. 5.6a shows the simulated average age, the average peak-age ($\Delta_{peak,1}$) and the lower bound on the average age (Δ_{LB}), as computed in the previous section for stream $\mathcal{U}_1$, and the average age ($\Delta_{\mathcal{U}_2}$) of stream $\mathcal{U}_2$. In this plot, we fix $\mu_1 = 10$, $\mu_2 = 5$, $\lambda_1 = 2$ and vary λ_2. As we can see, for stream $\mathcal{U}_1$ the average age, the lower bound, and the average peak-age grow without bounds when λ_2 gets close to $\frac{\mu_1\mu_2}{\lambda_1} - \mu_2$. This observation is in line with our result in Theorem 5.1 and the stability condition (5.1). In this simulation, we also notice that the average peak age and the lower bound appear to be good bounds on the average age, especially for small λ_2 and for values of $\lambda_2 \sim 0$ close to the limit $\frac{\mu_1\mu_2}{\lambda_1} - \mu_2$.

It is easy to see via a coupling argument that if we increase λ_2, the age process $\Delta_{\mathcal{U}_1}(t)$ of the $\mathcal{U}_1$ stream will stochastically increase. We see from the plots that the lower bound on $\Delta_{\mathcal{U}_1}$ and that its average peak-age exhibit the same behavior. However, the average age of stream $\mathcal{U}_2$ is decreasing in λ_2 (from (5.24)). Consequently, minimizing $\Delta_{\mathcal{U}_2}$ and minimizing $\Delta_{\mathcal{U}_1}$ are conflicting goals.

We have seen that the average age of stream $\mathcal{U}_2$ is not affected by the presence of the other stream. However, Fig. 5.6a shows the effect of stream $\mathcal{U}_2$ on the average age of stream $\mathcal{U}_1$ (Δ_1). For this, we plot the average age (Δ_{ref}) of an M/M/1 queue with generation rate $\lambda_1 = 2$ and service rate $\mu_1 = 10$ (given in [34]). We observe an expected behavior: for very low values of λ_2, the two average ages and the lower bound Δ_{LB} are close (they are all equal at $\lambda_2 = 0$). However, as λ_2 increases the presence of stream $\mathcal{U}_2$ quickly leads to an increase in Δ_1. In fact, for $\lambda_2 = 5$, Δ_1 is already 50% higher than Δ_{ref}. This shows that the presence of the priority stream $\mathcal{U}_2$ takes a heavy toll on the stream $\mathcal{U}_1$ age. Another observation is that the average age curve of stream $\mathcal{U}_2$ crosses the average age of stream $\mathcal{U}_1$ at a value of λ_2, denoted $\lambda_2^* = 1.9$. This means that for $\lambda_2 \leq \lambda_2^*$, stream $\mathcal{U}_2$ has an average age higher than stream $\mathcal{U}_1$. These observations show that not all values of λ_2 are suitable for our system. A small λ_2 will not ensure for stream $\mathcal{U}_2$ the priority it needs, whereas a

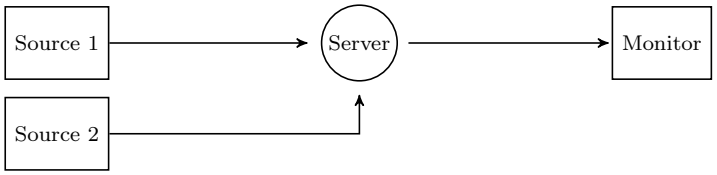


Figure 5.7 – Diagram representing the model with preemption for the low priority stream.

large λ_2 will make the average age of stream $\mathcal{U}_1$ large and the system unstable.

Fig. 5.6b plots the same quantities as Fig. 5.6a but under different settings: in this case, $\mu_1 = 6$, $\mu_2 = 18$, $\lambda_1 = 4$ and $\lambda_2 < 9$. In this particular scenario, we notice that the lower bound is a tight bound on the simulated average age for all values of λ_2, and it is tighter than the average peak age.

5.4 M/G/1/1 with Preemption for the Low-Priority Stream

Fig. 5.7 presents an illustration of the model. In this model, we assume we have no memory, hence packets from stream $\mathcal{U}_1$ preempt each other. However, if an arriving $\mathcal{U}_1$ packet finds the system busy serving a $\mathcal{U}_2$ packet, the server discards the stream $\mathcal{U}_1$ packet because stream $\mathcal{U}_2$ packets are given higher priority. Furthermore, the server applies a preemption policy whenever a packet from $\mathcal{U}_2$ is generated. This means that if a newly generated packet from stream $\mathcal{U}_2$ finds the system busy (serving a packet from $\mathcal{U}_1$ or $\mathcal{U}_2$), the server preempts the update currently in service and starts serving the new packet. Moreover, if the preempted packet belongs to $\mathcal{U}_1$ or $\mathcal{U}_2$, this packet is discarded.

These ideas are illustrated in part in Fig. 5.8, which also shows the variation of the instantaneous age of stream $\mathcal{U}_1$. In this plot, t_j refers to the generation time of the j^{th} packet, and D_i corresponds to the delivery time of the i^{th} successfully received packet of stream $\mathcal{U}_1$. As in this case not all the packets generated by source $\mathcal{U}_1$ are received, we distinguish between *generated* packets and *successful* packets. Moreover, t'_i and D'_i are the start and end times of the i^{th} period during which the system is busy serving packets only from stream $\mathcal{U}_2$.

5.4.1 Ages of Streams $\mathcal{U}_1$ and $\mathcal{U}_2$

Preliminaries

In this section also, unless stated otherwise, all random variables correspond to stream $\mathcal{U}_1$. We also follow the convention where a random variable U with no subscript corresponds to the steady-state version of U_j that refers to the random variable relative to the j^{th} received packet from stream $\mathcal{U}_1$. To differentiate between streams, we use superscripts, so that $U^{(i)}$ corresponds to the steady-state variable U relative to stream $\mathcal{U}_i$, $i = 1, 2$.

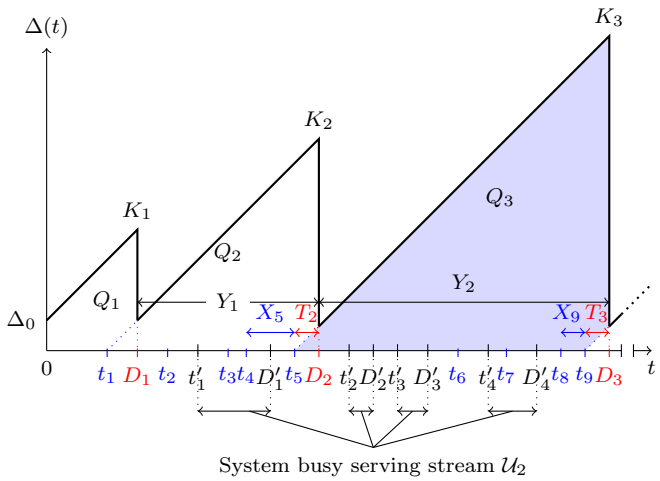


Figure 5.8 – Variation of the instantaneous age of stream $\mathcal{U}_1$.

In contrast to Section 5.3, here we follow exactly the notation introduced in Chapter 2, meaning

- $X^{(i)}$ is the interarrival time between two consecutive generated updates from stream $\mathcal{U}_i$, so $f_{X^{(i)}}(x) = \lambda_i e^{-\lambda_i x}$, $i = 1, 2$,
- $S^{(i)}$ is the service time random variable of stream $\mathcal{U}_i$ updates with p.d.f. $f_{S^{(i)}}(t)$, $i = 1, 2$,
- T_j is the system time, or the time spent by the j^{th} successfully received stream $\mathcal{U}_1$ update in the queue,
- Y_j is the interdeparture time between the j^{th} and $j + 1^{th}$ successfully received stream $\mathcal{U}_1$ updates.
- $R(\tau) = \max\{n : D_n \leq \tau\}$ is the number of successfully received updates from stream 1 in the interval $[0, \tau]$.

The reuse of the symbol Y_j, as it is defined in Chapter 2, is due to the fact that we use the interdeparture time approach (DTA) to compute the average age and average peak-age relative to source $\mathcal{U}_1$. Moreover, given that in this model there is no waiting in the queue, the system time of a received packet is equal to its service time. In our model, we assume the service time of the updates from the different streams to be independent of the interarrival time between consecutive packets (regardless if they belong to the same stream).

Finally, two important quantities that we will use extensively are

- $P_\lambda = \mathbb{E}\left(e^{-\lambda S^{(1)}}\right) = \int f_{S^{(1)}}(t) e^{-\lambda t} \mathrm{d}t$,
- $L_{\lambda_2} = \mathbb{E}\left(e^{-\lambda_2 S^{(2)}}\right) = \int f_{S^{(2)}}(t) e^{-\lambda_2 t} \mathrm{d}t$.

These are the Laplace transform of $f_{S^{(1)}}(t)$ and $f_{S^{(2)}}(t)$ evaluated at $\lambda = \lambda_1 + \lambda_2$ and λ_2, respectively.

Average Age and Average Peak age of Stream $\mathcal{U}_1$

Lemma 5.4. *For the priority preemption system described above, the moment generating function of the system time T corresponding to stream $\mathcal{U}_1$ is given by*

$$\phi_T(s) = \frac{P_{\lambda-s}}{P_\lambda}. \tag{5.25}$$

Note that the right-hand side of (5.25) *does not depend on the chosen stream.*

Proof. We use the same proof as in Lemma 4.2 with the number of streams $M = 2$. All variables in this proof corresponds to stream $\mathcal{U}_1$. The system time T_j of the j^{th} successfully received packet corresponds to the service time of the j^{th} received packet given that service was completed before any new arrival (because any new packet from any stream will preempt the current update being served). Therefore, in steady state, $\mathbb{P}(T > t) = \mathbb{P}\left(S^{(1)} > t | S^{(1)} < \min\left(X^{(1)}, X^{(2)}\right)\right)$. Hence, for $L = \min\left(X^{(1)}, X^{(2)}\right)$,

$$\begin{aligned} f_T(t) &= \lim_{\epsilon\to 0} \frac{\mathbb{P}(T \in [t, t+\epsilon])}{\epsilon} \\ &= \lim_{\epsilon\to 0} \frac{\mathbb{P}\left(S^{(1)} \in [t, t+\epsilon] | S^{(1)} < L\right)}{\epsilon} \\ &= \lim_{\epsilon\to 0} \frac{\mathbb{P}\left(S^{(1)} \in [t, t+\epsilon]\right) \mathbb{P}\left(S^{(1)} < L | S^{(1)} \in [t, t+\epsilon]\right)}{\epsilon \mathbb{P}\left(S^{(1)} < L\right)} \\ &= \frac{f_{S^{(1)}}(t)\mathbb{P}(L > t)}{\mathbb{P}\left(S^{(1)} < L\right)} = \frac{f_{S^{(1)}}(t)e^{-\lambda t}}{\mathbb{P}\left(S^{(1)} < L\right)}, \end{aligned}$$

where the last equality is due to the fact that L is exponentially distributed with rate $\lambda = \lambda_1 + \lambda_2$. Thus,

$$\phi_T(s) = \mathbb{E}\left(e^{sT}\right) = \int_0^\infty \frac{f_{S^{(1)}}(t)}{\mathbb{P}\left(S^{(1)} < L\right)} e^{-(\lambda-s)t} \mathrm{d}t = \frac{P_{\lambda-s}}{\mathbb{P}\left(S^{(1)} < L\right)}.$$

Finally,

$$\mathbb{P}\left(S^{(1)} < L\right) = \int_0^\infty f_{S^{(1)}}(t)\mathbb{P}(L > t)\,\mathrm{d}t = \int_0^\infty f_{S^{(1)}}(t)e^{-\lambda t}\mathrm{d}t = P_\lambda.$$

□

Lemma 5.5. *The moment generating function of the interdeparture time of stream $\mathcal{U}_1$, Y, is*

$$\phi_Y(s) = \frac{\lambda_1 P_{\lambda-s}\left(\lambda_2 L_{\lambda_2-s} - s\right)}{\lambda_1 P_{\lambda-s}\left(\lambda_2 L_{\lambda_2-s} - s\right) - s(\lambda_2 - s)}. \tag{5.26}$$

Proof. We use the detour flow graph method, introduced in Chapter 4. We define $\Lambda = \min\left(X^{(1)}, X^{(2)}\right)$. As Λ is the minimum of independent exponential random variables, then they are also exponentially distributed with rates $\lambda = \lambda_1 + \lambda_2$. Fig. 5.9

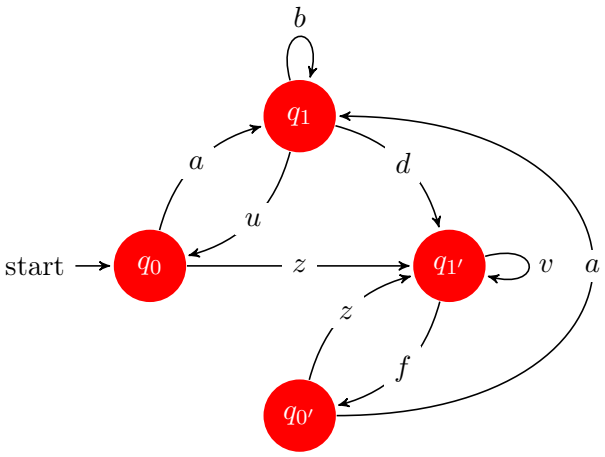


Figure 5.9 – Semi-Markov chain representing the M/G/1/1 interdeparture time for stream $\mathcal{U}_1$.

shows the semi-Markov chain relative to the interdeparture time Y_j between the j^{th} and $j+1^{th}$ successfully received packet of stream $\mathcal{U}_1$. When the j^{th} packet is delivered to the monitor, the system is in the idle state q_0. Due to the memoryless property of the interarrival times of both streams, two clocks start: a clock $X^{(1)}$ and a clock $X^{(2)}$. Clock $X^{(1)}$ ticks first with probability $a = \mathbb{P}\left(X^{(1)} < X^{(2)}\right)$, at which point a new packet from stream $\mathcal{U}_1$ will be generated first and the system goes to state q_1. The value A of the clock when it ticks has distribution $\mathbb{P}\left(A > t\right) = \mathbb{P}\left(X^{(1)} > t | X^{(1)} < X^{(2)}\right)$. Clock $X^{(2)}$ ticks first with probability $z = 1 - a = \mathbb{P}\left(X^{(2)} < X^{(1)}\right)$, at which point a new packet from stream $\mathcal{U}_2$ is generated first and the system goes to state $q_{1'}$. The value Z of this second clock when it ticks has distribution $\mathbb{P}\left(Z > t\right) = \mathbb{P}\left(X^{(2)} > t | X^{(2)} < X^{(1)}\right)$.

When the system arrives in state q_1, this means a packet from stream $\mathcal{U}_1$ is starting its service. Thus, due to the memoryless property of $X^{(2)}$, three clocks start: a service clock $S^{(1)}$, clock $X^{(1)}$ and clock $X^{(2)}$. The service clock ticks first with probability $u = \mathbb{P}\left(S^{(1)} < \Lambda\right)$ and its value U has distribution $\mathbb{P}\left(U > t\right) = \mathbb{P}\left(S^{(1)} > t | S^{(1)} < \Lambda\right)$. At this point, the stream $\mathcal{U}_1$ packet currently being served finishes service before any new packet is generated and the system goes back to state q_0. This ends the interdeparture time Y_j. Clock $X^{(1)}$ ticks first with probability $b = \mathbb{P}\left(X^{(1)} < \min\left(S^{(1)}, X^{(2)}\right)\right)$ and its value B has distribution $\mathbb{P}\left(B > t\right) = \mathbb{P}\left(X^{(1)} > t | X^{(1)} < \min\left(S^{(1)}, X^{(2)}\right)\right)$. At this point, a new stream $\mathcal{U}_1$ update is generated before any other update from other streams and preempts the one currently in service. In this case the system stays in state q_1. The third clock $X^{(2)}$ ticks first with probability $d = \mathbb{P}\left(X^{(2)} < \min\left(S^{(1)}, X^{(1)}\right)\right)$ and its value D has distribution $\mathbb{P}\left(D > t\right) = \mathbb{P}\left(X^{(2)} > t | X^{(2)} < \min\left(S^{(1)}, X^{(1)}\right)\right)$. At this point, a new update from stream $\mathcal{U}_2$ is generated, preempts the one currently in service and the system switches to state $q_{1'}$.

When the system arrives in state $q_{1'}$, this means a packet from stream $\mathcal{U}_2$ is starting its service. Thus, due to the memoryless property of $X^{(1)}$, two clocks are of interest: a service clock $S^{(2)}$ and clock $X^{(2)}$. What happens to stream $\mathcal{U}_1$ is irrelevant, as it has lower priority and any generated packet will be discarded. The service clock ticks

first with probability $f = \mathbb{P}\left(S^{(2)} < X^{(2)}\right)$ and its value F is distributed according to $\mathbb{P}\left(F > t\right) = \mathbb{P}\left(S^{(2)} > t | S^{(2)} < X^{(2)}\right)$. At this point, the stream $\mathcal{U}_2$ packet currently being served finishes service before any new packet is generated and the system goes to state $q_{0'}$. Otherwise, clock $X^{(2)}$ ticks first with probability $v = 1 - f = \mathbb{P}\left(X^{(2)} < S^{(2)}\right)$ and has value V distributed as $\mathbb{P}\left(V > t\right) = \mathbb{P}\left(X^{(2)} > t | X^{(2)} > S^{(2)}\right)$. At this point, a new update from stream $\mathcal{U}_2$ is generated, preempts the one currently in service and the system stays in state $q_{1'}$.

Finally, when the system arrives in state $q_{0'}$, this means the system is idle but no update from stream $\mathcal{U}_1$ has been delivered. Given that $X^{(1)}$ and $X^{(2)}$ are memoryless, the system in state $q_{0'}$ behaves exactly as if it were in state q_0.

From the above analysis, we see that the interdeparture time is given by the sum of the values of the different clocks on the path starting and finishing at q_0. For example, for the path $q_0 q_1 q_{1'} q_{0'} q_{1'} q_{0'} q_1 q_0$ in Fig. 5.9, the interdeparture time $Y = A_1 + D_1 + F_1 + Z_1 + F_2 + A_2 + U_1$, where all the random variables in the sum are mutually independent. This value of Y is also valid for the path $q_0 q_{1'} q_{0'} q_1 q_{1'} q_{0'} q_1 q_0$. Hence Y depends on the variables $A_j, B_j, D_j, F_j, U_j, V_j, Z_j$ and their number of occurrences and not on the path itself. Therefore, the probability that exactly $(i_1, i_2, i_3, i_4, i_5, i_6, i_7)$ occurrences of (A, B, D, F, U, V, Z) happen, which is equivalent to the probability that

$$Y = \sum_{k=1}^{i_1} A_k + \sum_{k=1}^{i_2} B_k + \sum_{k=1}^{i_3} D_k + \sum_{k=1}^{i_4} F_k + \sum_{k=1}^{i_5} U_k + \sum_{k=1}^{i_6} V_k + \sum_{k=1}^{i_7} Z_k,$$

is given by

$$a^{i_1} b^{i_2} d^{i_3} f^{i_4} u^{i_5} v^{i_6} z^{i_7} Q(i_1, i_2, i_3, i_4, i_5, i_6, i_7),$$

where $Q(i_1, i_2, i_3, i_4, i_5, i_6, i_7)$ is the number of paths with this combination of occurrences. Taking into account the fact that the $\{A_k, B_k, D_k, F_k, U_k, V_k, Z_k\}$ are mutually independent, the moment generating function of Y is

$$\begin{aligned}
\phi_Y(s) &= \mathbb{E}\left(\mathbb{E}\left(e^{sY} \,|\, (I_1, I_2, I_3, I_4, I_5, I_6, I_7) = (i_1, i_2, i_3, i_4, i_5, i_6, i_7)\right)\right) \\
&= \sum_{\substack{i_1, i_2, i_3, \\ i_4, i_5, i_6, i_7}} \Big[a^{i_1} b^{i_2} d^{i_3} f^{i_4} u^{i_5} v^{i_6} z^{i_7} Q(i_1, i_2, i_3, i_4, i_5, i_6, i_7) \\
&\quad \mathbb{E}\left(e^{s\left(\sum_{k=1}^{i_1} A_k + \sum_{k=1}^{i_2} B_k + \sum_{k=1}^{i_3} D_k + \sum_{k=1}^{i_4} F_k + \sum_{k=1}^{i_5} U_k + \sum_{k=1}^{i_6} V_k + \sum_{k=1}^{i_7} Z_k\right)}\right)\Big] \\
&= \sum_{\substack{i_1, i_2, i_3, \\ i_4, i_5, i_6, i_7}} \Big[a^{i_1} b^{i_2} d^{i_3} f^{i_4} u^{i_5} v^{i_6} z^{i_7} Q(i_1, i_2, i_3, i_4, i_5, i_6, i_7) \\
&\quad \mathbb{E}\left(e^{sA}\right)^{i_1} \mathbb{E}\left(e^{sB}\right)^{i_2} \mathbb{E}\left(e^{sD}\right)^{i_3} \mathbb{E}\left(e^{sF}\right)^{i_4} \mathbb{E}\left(e^{sU}\right)^{i_5} \mathbb{E}\left(e^{sV}\right)^{i_6} \mathbb{E}\left(e^{sZ}\right)^{i_7}\Big], \quad (5.27)
\end{aligned}$$

where $\{I_1, I_2, I_3, I_4, I_5, I_6, I_7\}$ are the random variables associated with the number of occurrences of $\{A, B, D, F, U, V, Z\}$, respectively.

However (5.27) is simply the generating function $H(W_1, W_2, W_3, W_4, W_5, W_6, W_7)$ of the detour flow graph shown in Fig. 5.10, where $W_1, W_2, W_3, W_4, W_5, W_6, W_7$ are

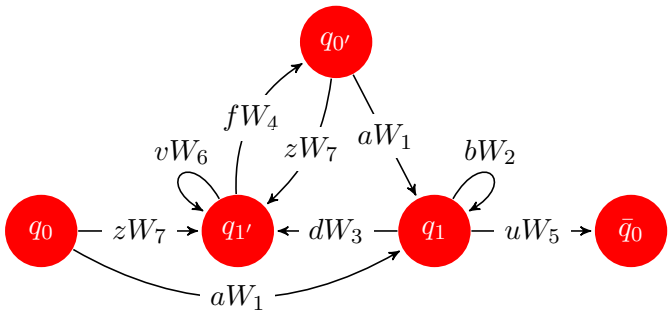


Figure 5.10 – Detour flow graph of the M/G/1/1 interdeparture time for stream $\mathcal{U}_1$.

dummy variables (see [56, pp. 213–216]). Observe that the sum in (5.27) is the transfer function from q_0 to $\bar{q}_0$ in the graph shown in Fig. 5.10 with

$$\begin{aligned}&(W_1, W_2, W_3, W_4, W_5, W_6, W_7)\\ &= \left(\mathbb{E}\left(e^{sA}\right), \mathbb{E}\left(e^{sB}\right), \mathbb{E}\left(e^{sD}\right), \mathbb{E}\left(e^{sF}\right), \mathbb{E}\left(e^{sU}\right), \mathbb{E}\left(e^{sV}\right), \mathbb{E}\left(e^{sZ}\right)\right).\end{aligned}$$

Solving the system of linear equations above yields the transfer function as

$$\begin{aligned}&H(W_1, W_2, W_3, W_4, W_5, W_6, W_7)\\ &= \sum_{\substack{i_1,i_2,i_3,\\ i_4,i_5,i_6,i_7}} \left[Q(i_1, i_2, i_3, i_4, i_5, i_6, i_7) a^{i_1} b^{i_2} d^{i_3} f^{i_4} u^{i_5} v^{i_6} z^{i_7}\right.\\ &\qquad \left. W_1^{i_1} W_2^{i_2} W_3^{i_3} W_4^{i_4} W_5^{i_5} W_6^{i_6} W_7^{i_7}\right]\\ &= \frac{uW_5 aW_1 (1 - vW_6)}{(1 - zW_7 fW_4 - vW_6)(1 - bW_2) - dW_3 aW_1 fW_4}. \end{aligned} \tag{5.28}$$

Thus

$$\phi_Y(s) = H\left(\mathbb{E}\left(e^{sA}\right), \mathbb{E}\left(e^{sB}\right), \mathbb{E}\left(e^{sD}\right), \mathbb{E}\left(e^{sF}\right), \mathbb{E}\left(e^{sU}\right), \mathbb{E}\left(e^{sV}\right), \mathbb{E}\left(e^{sZ}\right)\right).$$

Using Lemma 4.1 and Lemma 5.4, we know that

$$\begin{aligned}&\mathbb{E}\left(e^{sB}\right) = \mathbb{E}\left(e^{sD}\right) = \frac{\lambda(1 - P_{\lambda - s})}{(\lambda - s)(1 - P_\lambda)}, \quad \mathbb{E}\left(e^{sA}\right) = \mathbb{E}\left(e^{sZ}\right) = \frac{\lambda}{\lambda - s},\\ &\mathbb{E}\left(e^{sF}\right) = \frac{L_{\lambda_2 - s}}{L_{\lambda_2}} \quad \text{and} \quad \mathbb{E}\left(e^{sV}\right) = \frac{\lambda_2(1 - L_{\lambda_2 - s})}{(\lambda_2 - s)(1 - L_{\lambda_2})}.\end{aligned}$$

Moreover, we can notice that U has the same distribution as the system time T so $\mathbb{E}\left(e^{sU}\right) = \frac{P_{\lambda - s}}{P_\lambda}$. Simple computations show that

$$a = \frac{\lambda_1}{\lambda}, \quad b = \frac{\lambda_1}{\lambda}(1 - P_\lambda), \quad d = \frac{\lambda_2}{\lambda}(1 - P_\lambda), \quad f = L_{\lambda_2},$$

$$u = P_\lambda, \quad v = 1 - L_{\lambda_2}, \quad z = \frac{\lambda - \lambda_1}{\lambda}.$$

Finally, replacing the above expressions into (5.28), we get our result.

□

Theorem 5.4. *Assume an M/G/1/1 queue with preemption and a sender consisting of two sources generating packets according to two independent Poisson processes with rates λ_i, $i = 1, 2$, such that $\lambda = \lambda_1 + \lambda_2$. Moreover, packets belonging to stream i are served according to $S^{(i)}$. If stream $\mathcal{U}_2$ is given higher priority over stream $\mathcal{U}_1$, then*

1. *the average age of stream $\mathcal{U}_1$ is given by*

$$\Delta_1 = \frac{1}{\lambda_1 P_\lambda L_{\lambda_2}} + \frac{1 - L_{\lambda_2} - \lambda_2 \mathbb{E}\left(S^{(2)} e^{-\lambda_2 S^{(2)}}\right)}{\lambda_2 L_{\lambda_2}} \tag{5.29}$$

2. *and the average peak age of stream $\mathcal{U}_1$ is given by*

$$\Delta_{peak,1} = \frac{1}{\lambda_1 P_\lambda L_{\lambda_2}} + \frac{\mathbb{E}\left(S^{(1)} e^{-\lambda S^{(1)}}\right)}{P_\lambda}. \tag{5.30}$$

Proof. As in Chapters 3 and 4, to compute the average age, we use the DTA introduced in Section 2.3. We have shown that the average age for Stream 1 of the M/G/1/1 queue can be also expressed as the sum of the geometric areas Q_i under the instantaneous age curve of Fig. 5.8:

$$\Delta = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau \Delta(t) \mathrm{d}t = \frac{\mathbb{E}(Q)}{\mathbb{E}(Y)}, \tag{5.31}$$

where Y is the steady-state counterpart of Y_j, Q is the steady-state counterpart of Q_j and the second equality is justified by the fact that $(Y_j, T_j)_{j \geq 1}$ is stationary jointly second-order-ergodic. Using a similar argument as in the proof of Lemma 3.4 and corollary 3.1 and given that the interarrival time of all streams are memoryless, then the interdeparture times, Y_j and Y_{j+1}, between two consecutive received updates are i.i.d. Hence, $R(\tau)$ forms a renewal process and by [58],

$$\lim_{\tau \to \infty} \frac{R(\tau) - 1}{\tau} = \frac{1}{\mathbb{E}(Y)},$$

where Y is the steady-state interdeparture random variable. Introducing the quantity

$$C_j = \int_{D_j}^{D_{j+1}} \Delta(t) \mathrm{d}t$$

to be the reward function over the renewal period Y_j, we obtain using renewal reward theory [16, 58] that

$$\Delta = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau \Delta(t) \mathrm{d}t = \frac{\mathbb{E}(C_j)}{\mathbb{E}(Y_j)} = \frac{\mathbb{E}(Q_j)}{\mathbb{E}(Y_j)} < \infty.$$

This implies that $(T_j, Y_j)_{j \geq 1}$ is stationary jointly second-moment-ergodic.

Moreover, using Fig. 5.8, we see that, by applying the same argument presented in the proof of Lemma 3.4, the variables T_j and Y_j are independent for any $j \geq 1$. Thus, using (2.20),

$$\mathbb{E}\left(Q\right) = \frac{1}{2}\mathbb{E}\left(Y^2\right) + \mathbb{E}\left(TY\right) = \frac{1}{2}\mathbb{E}\left(Y^2\right) + \mathbb{E}\left(T\right)\mathbb{E}\left(Y\right).$$

Therefore,

$$\Delta_1 = \mathbb{E}(T) + \frac{\mathbb{E}(Y^2)}{2\mathbb{E}(Y)} \tag{5.32}$$

Moreover, from Fig. 5.8 we see that the peak age at the instant before receiving the j^{th} packet is given by

$$K_j = T_{j-1} + Y_{j-1}.$$

Hence, as given by (2.21), at steady state we get

$$\Delta_{peak,1} = \mathbb{E}(K) = \mathbb{E}(T) + \mathbb{E}(Y). \tag{5.33}$$

Using Lemma 5.4, we obtain $\mathbb{E}(T) = P_\lambda^{-1}\mathbb{E}\left(S^{(1)}e^{-\lambda S^{(1)}}\right)$. Using Lemma 5.5, we get that $\mathbb{E}(Y) = (\lambda_1 P_\lambda L_{\lambda_2})^{-1}$ and

$$\frac{\mathbb{E}(Y^2)}{2\mathbb{E}(Y)} = -\frac{1}{\lambda_2} - \frac{\mathbb{E}\left(S^{(1)}e^{-\lambda S^{(1)}}\right)}{P_\lambda} - \frac{\mathbb{E}\left(S^{(2)}e^{-\lambda_2 S^{(2)}}\right)}{L_{\lambda_2}} + \frac{1}{\lambda_1 P_\lambda L_{\lambda_2}} + \frac{1}{\lambda_2 L_{\lambda_2}}.$$

Using these expressions in (5.32) and (5.33), we achieve our result for stream $\mathcal{U}_1$. □

Average Age of Stream $\mathcal{U}_2$

By design, stream $\mathcal{U}_2$ is not at all interfered by stream $\mathcal{U}_1$ hence behaves like a traditional M/M/1/1 with preemption queue with generation rate λ_2 and service time $S^{(2)}$. The average age of this stream was computed in Chapter 4 to be

$$\Delta_2 = \frac{1}{\lambda_2 L_{\lambda_2}}. \tag{5.34}$$

5.5 Discussion

A close observation of Equations (5.29) and (5.30) leads to the following remarks:

- If the service time for stream $\mathcal{U}_2$ is 0, $\Delta_1 = \frac{1}{\lambda_1 P_\lambda} \geq \frac{1}{\lambda_1 P_{\lambda_1}}$, where $\Delta = \frac{1}{\lambda_1 P_{\lambda_1}}$ is the value of the average of stream $\mathcal{U}_1$ if stream $\mathcal{U}_2$ is not present. This result is due to the fact that whenever a stream $\mathcal{U}_2$ packet is generated, it immediately preempts the stream $\mathcal{U}_1$ packet being served hence increases the instantaneous age of the latter stream.

- By using L'Hopital's rule, we can show that

  $$\lim_{\lambda_2 \to 0} \Delta_1 = \Delta = \frac{1}{\lambda_1 P_{\lambda_1}}$$

  as it is expected. The average peak-age also converges to its value when no stream $\mathcal{U}_2$ exists.

- *Special case*: assume $S^{(1)} \sim \text{Exp}(\mu_1)$ and $S^{(2)} \sim \text{Exp}(\mu_2)$. Then

$$\Delta_1 = \frac{(\mu_1+\lambda_1)}{\lambda_1\mu_1}\left(\frac{\mu_2+\lambda_2}{\mu_2}\right) + \frac{\lambda_2}{\mu_2}\left(\frac{\mu_2+\lambda_2}{\lambda_1\mu_1} + \frac{1}{\mu_2+\lambda_2}\right) \tag{5.35}$$

and

$$\Delta_{peak,1} = \frac{1}{\mu_1+\lambda_1+\lambda_2} + \frac{(\mu_1+\lambda_1+\lambda_2)(\mu_2+\lambda_2)}{\lambda_1\mu_1\mu_2}. \tag{5.36}$$

Denoting $\Delta_{Norm} = \frac{\mu_1+\lambda_1}{\mu_1\lambda_1}$ to be the average age of stream $\mathcal{U}_1$ when stream $\mathcal{U}_2$ does not exist, we can compute the additional age the presence of stream $\mathcal{U}_2$ costs to stream $\mathcal{U}_1$:

$$\begin{aligned}\Delta_{diff} &= \Delta_1 - \Delta_{Norm} \\ &= \frac{\lambda_2}{\lambda_1\mu_1} + \frac{\lambda_2}{\lambda_1\mu_2} + \frac{\lambda_2}{\mu_1\mu_2} + \frac{\lambda_2^2}{\lambda_1\mu_1\mu_2} + \frac{\lambda_2}{\mu_2(\mu_2+\lambda_2)}.\end{aligned}$$

By letting $\mu_2 \to \infty$ we obtain $\Delta_{diff} \to \frac{\lambda_2}{\lambda_1\mu_1} > 0$, and by taking $\lambda_2 = 0$ we obtain $\Delta_{diff} = 0$ as predicted by the previous two remarks.

- Using (5.14), (5.35) and (5.36), we compare the performance of the preemption policy on stream $\mathcal{U}_1$ with that of the FCFS scheme from an age point of view when the service times corresponding to both sources are exponential. Fig. 5.11 plots the average ages and average peak-ages relative to stream $\mathcal{U}_1$ for the preemption, as well as for the FCFS schemes. In both cases, we assume stream $\mathcal{U}_1$ packets are generated according to a Poisson process of rate $\lambda_1 = 2$ and served according to an exponential service time with rate $\mu_1 = 10$. As for stream $\mathcal{U}_2$ updates, they are generated according to a Poisson process with rate λ_2 and served according to an exponential service time with rate $\mu_2 = 5$. We observe from Fig. 5.11 that the preemption scheme performs worse than the FCFS except when λ_2 is close to the FCFS stability condition. This observation comes as a surprise because we would think that the constraint of delivering all generated packets imposed by a FCFS system would pull the age up, compared to the more flexible preemptive scheme. However, we can explain this result in the following way: When using the preemptive scheme and not storing any updates, the system incurs a substantial idle time (from the source $\mathcal{U}_1$ point of view) during which it waits for a new stream $\mathcal{U}_1$ update to be generated. In fact, this is a direct consequence of the first remark. Interestingly, Bedewy et al. in [6] show that for a single source and exponential service time, the optimal policy to adopt is the preemptive scheme. Fig. 5.11 proves that the introduction of an another source with higher priority has a significant impact on the performances of the different transmission schemes. For instance, the preemption scheme is not optimal anymore even for exponential service times.

5.6 Conclusion

In this chapter, we have studied the effect of implementing content-dependent policies on the average age of the packets. We have considered a sender that generates two independent Poisson streams with one stream having higher priority than the other

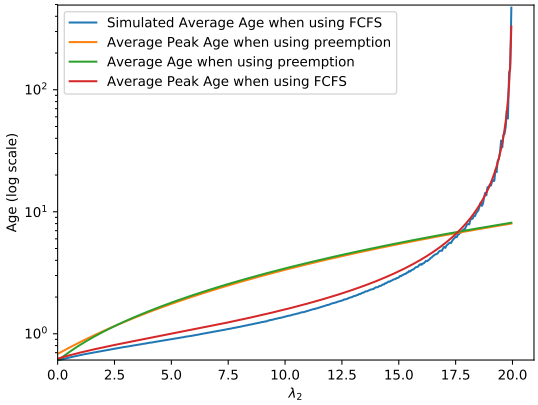


Figure 5.11 – Comparison between the average peak ages of the low priority source $\mathcal{U}_1$ when using the FCFS and the preemption schemes and exponential service times. We fix $\lambda_1 = 2$, $\mu_1 = 10$, $\mu_2 = 5$.

stream. The "high priority" stream is sent using a preemption policy, whereas at first the "regular" stream is transmitted using a FCFS policy and then it is transmitted using preemption. We derived the stability condition for the former system, as well as closed-form expressions for the average peak-age and a lower bound on the average age of the "regular" stream. For the latter system we have given exact expressions for the average age and average peak-age relative to the "regular" stream and we have shown through simulations that, even if the service times relative to both streams are exponential, preemption is not the optimal strategy to adopt for the "regular" stream. In fact, for fixed service rates and "regular" stream generation rate, the FCFS strategy performs better for a large interval of "high priority"- stream generation rate.

5.7 Appendix

5.7.1 Proof of Theorem 5.1

Assume that

$$\mu_1 > \lambda_1 \left(1 + \frac{\lambda_2}{\mu_2}\right). \tag{5.37}$$

The detailed balance equations of the Markov chain given by Fig. 5.3 are given by:

$$\begin{cases} \lambda\pi_0 = \mu_1\pi_1 + \mu_2\pi_1', \\ (\lambda_1 + \mu_2)\pi_1' = \lambda_2\pi_0, \\ \text{for } i \geq 1, \\ \pi_{i+1} = \left(1 + \dfrac{\lambda}{\mu_1} - \dfrac{\mu_2\lambda_2}{\mu_1(\mu_2 + \lambda_1)}\right)\pi_i - \dfrac{\mu_2\lambda_1}{\mu_1(\mu_2 + \lambda_1)}\pi_i' \\ \qquad\qquad - \dfrac{\lambda_1}{\mu_1}\pi_{i-1}, \\ \pi_{i+1}' = \dfrac{\lambda_2}{\mu_2 + \lambda_1}\pi_i + \dfrac{\lambda_1}{\mu_2 + \lambda_1}\pi_i', \end{cases} \tag{5.38}$$

where $\lambda = \lambda_1 + \lambda_2$. For easier notation we denote

$$a_1 = 1 + \frac{\lambda}{\mu_1} - \frac{\mu_2\lambda_2}{\mu_1(\mu_2+\lambda_1)},$$
$$a_2 = \frac{\mu_2\lambda_1}{\mu_1(\mu_2+\lambda_1)},$$
$$a_3 = \frac{\lambda_1}{\mu_1},$$
$$a_4 = \frac{\lambda_2}{\mu_2+\lambda_1},$$
$$a_5 = \frac{\lambda_1}{\mu_2+\lambda_1}.$$

Rewriting (5.38) in matrix form and using the above notation, we get

$$\begin{bmatrix} \pi_{i+1} \\ \pi'_{i+1} \\ \pi_i \\ \pi'_i \end{bmatrix} = \begin{bmatrix} a_1 & -a_2 & -a_3 & 0 \\ a_4 & a_5 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \pi_i \\ \pi'_i \\ \pi_{i-1} \\ \pi'_{i-1} \end{bmatrix}.$$

Let $\mathbf{A}_i = \begin{bmatrix} \pi_{i+1} \\ \pi'_{i+1} \\ \pi_i \\ \pi'_i \end{bmatrix}$, $\mathbf{C} = \begin{bmatrix} a_1 & -a_2 \\ a_4 & a_5 \end{bmatrix}$, $\mathbf{D} = \begin{bmatrix} -a_3 & 0 \\ 0 & 0 \end{bmatrix}$ and $\mathbf{H} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{I}_2 & \mathbf{0} \end{bmatrix}$. Then

$$\mathbf{A}_i = \mathbf{H}\mathbf{A}_{i-1}.$$

Thus

$$\mathbf{A}_i = \mathbf{H}^i\mathbf{A}_0,\ i \geq 0 \tag{5.39}$$

where $\mathbf{A}_0 = \begin{bmatrix} \pi_1 \\ \pi'_1 \\ \pi_0 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{\lambda}{\mu_1} - \frac{\mu_2\lambda_2}{\mu_1(\lambda_1+\mu_2)} \\ \frac{\lambda_2}{\lambda_1+\mu_2} \\ 1 \\ 0 \end{bmatrix} \pi_0$, using the first two equations of system (5.38).

(5.39) shows that in order to find the stability criterion of the system in (5.38) we first need to study the properties of $\mathbf{H}$. For that we compute its eigenvalues l_0, l_1, l_2, l_3 by solving the characteristic equation $|l\mathbf{I}_4 - \mathbf{H}| = 0$. This leads to

$$|l\mathbf{I}_4 - \mathbf{H}| = l(l-1)(l^2 - l(a_1 + a_5 - 1) + a_3a_5). \tag{5.40}$$

$\mathbf{H}$ has two obvious eigenvalues $l_0 = 0$ and $l_3 = 1$. To find the last two eigenvalues, let's find the root of the quadratic polynomial

$$p(l) = l^2 - l(a_1 + a_5 - 1) + a_3a_5. \tag{5.41}$$

It can be shown through simple algebra that the discriminant of the above polynomial is strictly positive. Hence the remaining eigenvalues l_1 and l_2 are real and distinct. Let's assume that $l_1 < l_2$. This means that the matrix $\mathbf{H}$ is diagonalizable and can be written as

$$\mathbf{H} = \mathbf{B}\mathbf{\Lambda}\mathbf{B}^{-1},$$

where the columns of $\mathbf{B}$ are the eigenvectors of $\mathbf{H}$ and form a basis of $\mathbb{R}^4$. We denote by $\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ the eigenvectors corresponding to l_0, l_1, l_2, l_3.

So we can write $\mathbf{A}_0$ as

$$\mathbf{A}_0 = (\alpha_0\mathbf{e}_0 + \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \alpha_3\mathbf{e}_3)\pi_0, \tag{5.42}$$

with $\alpha_0, \alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$. Hence for $i > 0$,

$$\begin{aligned}\mathbf{A}_i &= \mathbf{H}^i\mathbf{A}_0 \\ &= (\alpha_0\mathbf{H}^i\mathbf{e}_0 + \alpha_1\mathbf{H}^i\mathbf{e}_1 + \alpha_2\mathbf{H}^i\mathbf{e}_2 + \alpha_3\mathbf{H}^i\mathbf{e}_3)\pi_0 \\ &= (\alpha_0 l_0^i\mathbf{e}_0 + \alpha_1 l_1^i\mathbf{e}_1 + \alpha_2 l_2^i\mathbf{e}_2 + \alpha_3 l_3^i\mathbf{e}_3)\pi_0 \\ &= (\alpha_1 l_1^i\mathbf{e}_1 + \alpha_2 l_2^i\mathbf{e}_2 + \alpha_3\mathbf{e}_3)\pi_0, \end{aligned} \tag{5.43}$$

since $l_0 = 0$ and $l_3 = 1$. Equation (5.43) shows that three conditions need to be satisfied for the system to be stable and a steady-state distribution to exist:

- **Condition 1:** $|l_1| < 1$ and $|l_2| < 1$.
- **Condition 2:** $\alpha_3 = 0$.
- **Condition 3:** $\alpha_1 l_1^i\mathbf{e}_1 + \alpha_2 l_2^i\mathbf{e}_2$ has positive components for all $i > 0$.

Condition 1 and **Condition 2** ensure that

$$\lim_{i\to\infty} \pi_i = \lim_{i\to\infty} \pi_i' = 0.$$

Condition 3 makes sure that the components of $\mathbf{A}_i$ are positive probabilities. We will show that (5.37) is sufficient for the above three conditions to hold.

Given that l_1 and l_2 are the roots of (5.41) then the following holds

$$\begin{aligned} l_1 l_2 &= a_3 a_5 \\ l_1 + l_2 &= a_1 + a_5 - 1. \end{aligned} \tag{5.44}$$

However, $l_1 l_2 = a_3 a_5 = \frac{\lambda_1^2}{\mu_1(\mu_2+\lambda_1)} \geq 0$. This means that either both l_1 and l_2 are positive or they are both negative. Using (5.44) again, we notice that

$$l_1 + l_2 = a_1 + a_5 - 1 = \frac{\lambda_1\mu_2 + \lambda_1^2 + \lambda_1\lambda_2 + \lambda_1\mu_1}{\mu_1(\mu_2 + \lambda_1)} \geq 0.$$

This shows that both l_1 and l_2 are strictly positive (since 0 is not a root of $p(l)$). So to prove that **Condition 1** is satisfied we need to prove that $l_1 < l_2 < 1$. This is equivalent to show that (5.41) evaluated at 1 is strictly positive and that $l_1 l_2 < 1$ since $p(l)$ is a convex quadratic function in $l > 0$. Using simple algebra it can be shown that

$$p(1) = 1 - (a_1 + a_5 - 1) + a_3 a_5 = \frac{\mu_1\mu_2 - \lambda_1(\mu_2 + \lambda_2)}{\mu_1(\mu_2 + \lambda_1)} > 0,$$

where the last inequality is due to (5.37). Moreover, (5.37) tells us that μ_1 should be strictly bigger that λ_1. Thus we get that

$$l_1 l_2 = \frac{\lambda_1}{\mu_1}\frac{\lambda_1}{\mu_2 + \lambda_1} < 1.$$

This shows that $0 < l_1 < l_2 < 1$ and that **Condition 1** is satisfied.

To prove **Condition 2** we start by computing the eigenvectors of $\mathbf{H}$. For $l_0 = 0$, we solve the system given by $\mathbf{H}\mathbf{e}_0 = \mathbf{0}$. If $\mathbf{e}_0 = \begin{bmatrix} u_1 & u_2 & u_3 & 1 \end{bmatrix}^T$ then

$$\begin{bmatrix} a_1 & -a_2 & -a_3 & 0 \\ a_4 & a_5 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

This system leads to $\mathbf{e}_0 = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^T$. Similarly, for $j = 1, 2, 3$, if $\mathbf{e}_j = \begin{bmatrix} u_1 & u_2 & u_3 & 1 \end{bmatrix}^T$ then solving the system

$$\begin{bmatrix} a_1 & -a_2 & -a_3 & 0 \\ a_4 & a_5 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ 1 \end{bmatrix} = l_j \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ 1 \end{bmatrix}$$

leads to $\mathbf{e}_j = \begin{bmatrix} l_j(l_j - a_5) & l_j a_4 & l_j - a_5 & a_4 \end{bmatrix}^T$.

We know that $\mathbf{H} = \mathbf{B}\mathbf{\Lambda}\mathbf{B}^{-1}$. If

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & l_2 & 0 & 0 \\ 0 & 0 & l_1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

then

$$\mathbf{B} = \begin{bmatrix} 1 - a_5 & l_2(l_2 - a_5) & l_1(l_1 - a_5) & 0 \\ a_4 & l_2 a_4 & l_1 a_4 & 0 \\ 1 - a_5 & l_2 - a_5 & l_1 - a_5 & 0 \\ a_4 & a_4 & a_4 & 1 \end{bmatrix}.$$

Note that the determinant of $\mathbf{B}$, $|\mathbf{B}|$, is non-zero when we assume (5.37). Indeed,

$$|\mathbf{B}| = a_4 a_5 (l_2 - l_1)(-2 + a_5 - a_3 a_5 + a_1) < 0$$

since $l_2 > l_1$ and $-2 + a_5 - a_3 a_5 + a_1 = -p(1) < 0$ as shown before. In order to compute α_3, we rewrite (5.42) as follows

$$\mathbf{A}_0 = \begin{bmatrix} \mathbf{e}_3 & \mathbf{e}_2 & \mathbf{e}_1 & \mathbf{e}_0 \end{bmatrix} \begin{bmatrix} \alpha_3 \\ \alpha_2 \\ \alpha_1 \\ \alpha_0 \end{bmatrix} \pi_0 = \mathbf{B} \begin{bmatrix} \alpha_3 \\ \alpha_2 \\ \alpha_1 \\ \alpha_0 \end{bmatrix} \pi_0.$$

But we also know that

$$\mathbf{A}_0 = \begin{bmatrix} \frac{\lambda}{\mu_1} - \frac{\mu_2\lambda_2}{\mu_1(\lambda_1+\mu_2)} \\ \frac{\lambda_2}{\lambda_1+\mu_2} \\ 1 \\ 0 \end{bmatrix} \pi_0 = \begin{bmatrix} a_1 - 1 \\ a_4 \\ 1 \\ 0 \end{bmatrix} \pi_0.$$

Thus

$$\mathbf{B} \begin{bmatrix} \alpha_3 \\ \alpha_2 \\ \alpha_1 \\ \alpha_0 \end{bmatrix} = \begin{bmatrix} a_1 - 1 \\ a_4 \\ 1 \\ 0 \end{bmatrix}. \tag{5.45}$$

Solving the system in (5.45) with respect to α_3, α_2, α_1 and α_0 we get that

$$\begin{bmatrix} \alpha_3 \\ \alpha_2 \\ \alpha_1 \\ \alpha_0 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{1}{l_2-l_1} \\ \frac{-1}{l_2-l_1} \\ 0 \end{bmatrix}.$$

Thus $\alpha_3 = 0$ and **Condition 2** is proved. Note that we didn't need any assumptions to prove this condition.

Given the above results, we can now rewrite the system in (5.43) as

$$\begin{cases} \mathbf{A}_i = (\alpha_2 l_2^i \mathbf{e}_2 + \alpha_1 l_1^i \mathbf{e}_1)\pi_0, \qquad i > 0 \\ \\ \mathbf{A}_0 = (\alpha_2 \mathbf{e}_2 + \alpha_1 \mathbf{e}_1)\pi_0 = \begin{bmatrix} a_1 - 1 \\ a_4 \\ 1 \\ 0 \end{bmatrix} \pi_0. \end{cases} \tag{5.46}$$

Using (5.46) we can prove **Condition 3**. In fact, for any $i > 0$,

$$\begin{aligned} \alpha_2 l_2^i \mathbf{e}_2 + \alpha_1 l_1^i \mathbf{e}_1 &\overset{(a)}{=} \alpha_2 \left(l_2^i \mathbf{e}_2 - l_1^i \mathbf{e}_1 \right) \\ &\overset{(b)}{\succ} \alpha_2 l_1^i (\mathbf{e}_2 - \mathbf{e}_1) \\ &\overset{(c)}{=} l_1^i \begin{bmatrix} a_1 - 1 \\ a_4 \\ 1 \\ 0 \end{bmatrix} \\ &\overset{(d)}{\succ} \mathbf{0}, \end{aligned}$$

where $\mathbf{x} \succ \mathbf{y}$ for some vectors $\mathbf{x}$ and $\mathbf{y}$ means that the components of $\mathbf{x} - \mathbf{y}$ are strictly positive and

- (a) is because $\alpha_2 = -\alpha_1$,
- (b) is because $0 < l_1 < l_2$,
- (c) is obtained from the second equality in (5.46),

(d) follows since $a_1 - 1 > 0$ and $a_4 > 0$.

Up till now we have shown that if $\mu_1 > \lambda_1 \left(1 + \frac{\lambda_2}{\mu_2}\right)$, the system described in Section 5.2 is stable and a steady-state distribution exists given by (5.46). The final point to prove in Theorem 5.1 is the expression of π_0. For that we solve for π_0 the following equation

$$\pi_0 + \sum_{i=1}^{\infty} \pi_i + \pi'_i = \pi_0 + \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix} \sum_{i=1}^{\infty} \mathbf{A}_i = 1.$$

Using the first equation of (5.46) and replacing α_1 and α_2 by their expressions in function of l_1 and l_2, using the fact that $l_1 + l_2$ and $l_1 l_2$ are given by (5.44) and finally replacing a_1, a_2, a_3, a_4 and a_5 by their expressions in function of λ_1, λ_2, μ_1, μ_2 we get

$$\pi_0 = \frac{\mu_2}{\mu_2 + \lambda_2} - \frac{\lambda_1}{\mu_1}.$$

5.7.2 Proof of Corollary 5.1

At any point in time, there are exactly i stream $\mathcal{U}_1$ packets in the system if we are in state q_i or q'_{i+1} in the Markov chain given by Fig. 5.3. This means that the probability of having exactly i stream $\mathcal{U}_1$ packets in the system is $\pi_i + \pi'_{i+1}$. Hence, using the same quantities as in Section 5.7.1

$$\begin{aligned}
\phi_{N(t)}(s) = \mathbb{E}\left(e^{sN(t)}\right) &= \sum_{n=0}^{\infty} e^{sn}(\pi_i + \pi'_{i+1}) = \sum_{n=0}^{\infty} e^{sn} \left(\mathbf{A}_n^T \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \right) \\
&= \sum_{n=0}^{\infty} e^{sn} \alpha_2 \pi_0 \left((l_2^n \mathbf{e}_2 - l_1^n \mathbf{e}_1)^T \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \right) \\
&= \alpha_2 \pi_0 \left(\sum_{n=0}^{\infty} (e^s l_2)^n \, \mathbf{e}_2^T \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} - \sum_{n=0}^{\infty} (e^s l_1)^n \, \mathbf{e}_1^T \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \right) \\
&= \alpha_2 \pi_0 \left(\frac{1}{1 - l_2 e^s}(l_2 a_4 + l_2 - a_5) - \frac{1}{1 - l_1 e^s}(l_1 a_4 + l_1 - a_5) \right) \\
&= \alpha_2 \pi_0 (l_2 - l_1) \left(\frac{a_4 + 1 - a_5 e^s}{1 - (l_1 + l_2)e^s + l_1 l_2 e^{2s}} \right). \qquad (5.47)
\end{aligned}$$

where the quantities used here are the one defined in the proof of Theorem 5.1. Thus,

$$\phi_{N(t)}(s) = \pi_0 \left(\frac{\mu_1 \left(\lambda_1 + \lambda_2 + \mu_2 - \lambda_1 e^s\right)}{\mu_1 \mu_2 + \mu_1 \lambda_1 - e^s(\lambda_1 \mu_2 + \lambda_1^2 + \lambda_1 \lambda_2 + \lambda_1 \mu_1) + \lambda_1^2 e^{2s}} \right).$$

This last equality is obtained by using (5.44), $\alpha_2 = \frac{1}{l_2 - l_1}$ and replacing a_1, a_2, a_3, a_4, a_5 by their expressions in function of λ_1, λ_2, μ_1 and μ_2 in (5.47). Finally,

$$\mathbb{E}(N(t)) = \left. \frac{\mathrm{d}\phi_{N(t)}(s)}{\mathrm{d}s} \right|_{s=0} = \frac{\lambda_1 \left(2\lambda_2 \mu_2 + \lambda_2 \mu_1 + \lambda_2^2 + \mu_2^2\right)}{(\mu_2 + \lambda_2)\left(\mu_1 \mu_2 - \lambda_1 (\mu_2 + \lambda_2)\right)}.$$

Part II

Age in the Presence of Noise

Status Updates through M/G/1/1 Queues with HARQ

6

6.1 Introduction

In the first part of this thesis, we assumed the channel to be error-free and always correctly delivering the transmitted packets. In this second part of the dissertation, we consider a more practical channel example: The erasure channel defined in Definition 1.2. This type of channel could model the Internet where packets can be lost (or erased) with a certain probability $\epsilon > 0$.

In this chapter[1], we assume updates are generated by a single source according to a Poisson process with rate λ. However, the system can handle only one update at a time without any buffer to store incoming updates. This means that whenever a new update is generated and the system is busy, the transmitter has to make a decision: does it give higher priority to the new update or to the one being transmitted? In other words, does it preempt or not? The two transmission schemes studied here are M/G/1/1 with blocking and M/G/1/1 with preemption (see Section 1.2.3). It has been shown that for exponential update service times, preemption ensures the lowest average age [35]. However, our results in Chapter 3 suggest that under the assumption of gamma distributed service time, preemption might not be the best policy.

This chapter answers the previous question when we assume updates are sent through a symbol erasure channel with erasure rate ϵ, while using hybrid ARQ (HARQ) protocols to combat erasures. Two HARQ protocols, introduced in [20], are studied: (a) infinite incremental redundancy (IIR) and (b) fixed redundancy (FR). In both cases we assume a generated update contains K information symbols. In IIR, encoding is performed at the physical layer where the K information symbols are encoded using a rateless code. Hence, the transmission of an update continues until $k_s = K$ unerased symbols are received. As for the FR, coding is applied at the physical and packet layer. This means that the update is divided into k_p packets with each packet

[1]The material in this chapter is based on [47, 48].

encoded using an (n_s, k_s)-Maximum Distance Separable (MDS) code. So, in this case, the total number of information symbols is $K = k_p k_s$. At the packet level we apply a rateless code and thus the transmission of an update terminates when k_p unerased packets are received. In order to decode a packet, the receiver needs to wait for n_s encoded symbols. Once received, a packet is declared erased if fewer than k_s symbols are successful. It is worth noting that in this setup we send one symbol per channel use and thus the arrival rate λ is the number of updates generated per channel use. The effect of these schemes on the transmission time of data was studied in [20]. It was shown that FR leads to a slower delivery than IIR. While the main aim of [20] is the successful delivery of every update, in this chapter we are ready to sacrifice some updates for fresher information.

This chapter is organized as follows: We first begin by deriving in Section 6.3 an expression for the average age under general service time distribution when we choose M/G/1/1 with blocking. In Section 6.4, we use this expression to compute the average age when we consider the IIR and FR protocols. Sections 6.5 and 6.6 follow the same logic but in this case we choose M/G/1/1 with preemption. Finally, Section 6.7 compares the performances of both models for a given HARQ protocol as well as the performance of both protocols given a model. We show that no matter the protocol, prioritizing the current update is better than preempting it. Moreover, in the case of FR, we show that no matter the model and for a fixed arrival rate λ, there exists an optimal codeword length n_s.

6.2 Preliminaries

In this chapter, we use the notation introduced in Chapter 2. This means that we call the updates that are not dropped, and thus delivered to the receiver, as "successfully received updates" or "successful updates". In addition to that, we also define:

- I_i to be the true index of the i^{th} successfully received update,
- t_j to be the generation time of the j^{th} packet (not necessarily successful),
- t'_{I_i} to be the reception time of the i^{th} successful packet,
- $Y_i = t'_{I_{i+1}} - t'_{I_i}$ to be the interdeparture time between two consecutive successfully received updates,
- $X_i = t_{I_i+1} - t_{I_i}$ to be the interarrival time between the successfully transmitted update and the next generated one (which may or may not be successfully transmitted), so $f_X(x) = \lambda e^{-\lambda x}$,
- S_j to be the service time of the j^{th} generated update with cumulative distribution probability $F_S(t)$. In the case of the M/G/1/1 with blocking scenario, S_i denotes the service time of the i^{th} successful packet,
- T_i to be the system time, or the time spent by the i^{th} successful update in the queue,

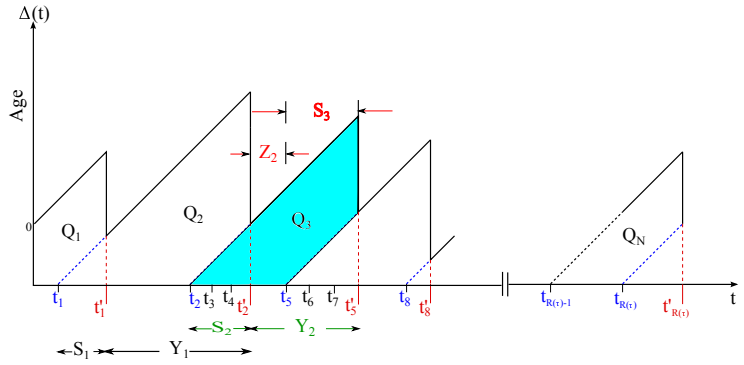


Figure 6.1 – Variation of the instantaneous age for M/G/1/1 with blocking

- $R(\tau) = \max\left\{n : t'_{I_n} \leq \tau\right\}$, to be the number of successfully received updates in the interval $[0, \tau]$.

In our models, we assume the service time S_k of the k^{th} update is independent from the interarrival time random variables $\{X_1, X_2, ..., X_k, ...\}$ and that the sequence $\{S_1, S_2, ...\}$ forms an i.i.d process.

From (1.2), Fig. 6.1 and Fig. 6.8, the average age for both M/G/1/1 queues can be also expressed as the sum of the geometric areas Q_i under the instantaneous age curve. We will use DTA (see Section 2.3) to compute the average age. Thus, as we have already shown in Chapter 3

$$\Delta = \lim_{\tau\to\infty} \frac{R(\tau)}{\tau} \frac{1}{R(\tau)} \sum_{i=1}^{R(\tau)} Q_i = \lambda_e \mathbb{E}(Q_i), \tag{6.1}$$

where $\lambda_e = \lim_{\tau\to\infty} \frac{R(\tau)}{\tau}$ and the second equality is due to the fact that $(Y_i, T_i)_{i\geq 1}$ is jointly second-moment-ergodic as we will see later.

6.3 M/G/1/1 with Blocking

In this setup, a generated update is discarded if it finds the system busy. This means an update is served only if it arrives at an idle system. This concept is depicted in Fig. 6.1: For instance, the update generated at time t_2 is served since the system is empty at that time. However, the updates generated at times t_3 and t_4 find the system busy and are thus discarded. One important note here is that the system time T_i of the i^{th} successful update is equal to its service time.

6.3.1 Average Age Calculation

Lemma 6.1. *For an M/G/1/1 blocking system we have,*

$$\lambda_e = \frac{1}{\mathbb{E}(Y)} = \frac{1}{\mathbb{E}(X) + \mathbb{E}(S)}, \tag{6.2}$$

where Y, X and S are the steady-state counterparts of the variables defined in Section 6.2.

Proof. $R(\tau)$ is a renewal process with inter-renewal time between two renewals given by the random variable Y. As shown in Fig. 6.1, the renewal period is the interval:

$$Y_i = Z_i + S_{i+1}. \tag{6.3}$$

Because each departure leaves the system empty and the interarrival times are memoryless, then the interval Z_i, which is the residual interarrival time until a new update is generated, is independent of Y_{i-1} and it has an exponential distribution. Hence, all the Y_i's are identically distributed and the Z_i's are stochastically equal to the interarrival time X. This proves why $R(\tau)$ is a renewal process. The claim follows [58]. □

Now we can compute the average age which is given by the following theorem,

Theorem 6.1. *The process $(Y_i, T_i)_{i\geq 1}$ is jointly second-moment-ergodic and the average age of an M/G/1/1 system with blocking is*

$$\Delta = \mathbb{E}(S)\left(\frac{\beta}{2}(C_S+1)+\frac{1}{\beta}\right), \tag{6.4}$$

where $C_S = \frac{Var(S)}{\mathbb{E}(S)^2}$ is the squared coefficient of variation and $\beta = \frac{\rho}{\rho+1}$ with $\rho = \frac{\mathbb{E}(S)}{\mathbb{E}(X)} = \lambda\mathbb{E}(S)$.

Proof. We have seen in Lemma 6.1 that $R(\tau)$ is a renewal process with $(Y_i)_{i\geq 1}$ being the inter-renewal intervals. By defining $D_i = \int_{t'_{I_i}}^{t'_{I_{i+1}}} \Delta(t)\mathrm{d}t$ to be the reward function over the renewal period Y_i, we get using renewal reward theory [16, 58] that

$$\Delta = \lim_{\tau\to\infty}\frac{1}{\tau}\int_0^\tau \Delta(t)\mathrm{d}t = \frac{\mathbb{E}(D_i)}{\mathbb{E}(Y_i)} = \frac{\mathbb{E}(Q_i)}{\mathbb{E}(Y_i)} < \infty.$$

This implies that $(T_i, Y_i)_{j\geq 1}$ is stationary jointly second-moment-ergodic. We need to compute the average area of the trapezoid Q_i. To do that, notice first that, using a similar argument as the one used in the proof of Lemma 6.1, the service time S_i and Y_i are independent. Thus,

$$\begin{aligned}\mathbb{E}(Q_i) &= \mathbb{E}\left(\frac{(S_{i-1}+Y_{i-1})^2}{2} - \frac{S_i^2}{2}\right)\\ &= \frac{1}{2}\mathbb{E}\left(Y_{i-1}^2\right) + \mathbb{E}(S_{i-1})\mathbb{E}(Y_{i-1}).\end{aligned} \tag{6.5}$$

Since we are interested in the steady-state behavior, we will drop the subscript index on the random variables. Hence,

$$\begin{aligned}\mathbb{E}(Q) &= \frac{1}{2}\mathbb{E}\left(Y^2\right) + \mathbb{E}(S)\mathbb{E}(Y)\\ &= \frac{1}{2}\mathbb{E}\left((X+S)^2\right) + \mathbb{E}(S)\mathbb{E}(S+X)\\ &= \frac{1}{2}\left(\mathbb{E}\left(X^2\right) + \mathbb{E}\left(S\right)^2\right) + \frac{1}{2}\mathrm{Var}(S) + 2\mathbb{E}(S)\mathbb{E}(X) + \mathbb{E}(S)^2\\ &= \frac{1}{2}\left(\mathbb{E}\left(S\right)^2 + \mathrm{Var}(S)\right) + \mathbb{E}\left(X\right)^2 + 2\mathbb{E}(S)\mathbb{E}(X) + \mathbb{E}(S)^2\\ &= \left(\mathbb{E}(X)+\mathbb{E}(S)\right)^2 + \frac{1}{2}\left(\mathbb{E}(S)^2 + \mathrm{Var}(S)\right),\end{aligned} \tag{6.6}$$

where the third equality is obtained by adding and subtracting $\frac{1}{2}\mathbb{E}(S)^2$ to the second equality, and the fourth equality is obtained by noticing that for the exponential random variable X we have $\mathbb{E}\left(X^2\right) = 2\mathbb{E}(X)^2$. Using (6.2) and (6.6), we get (6.4). □

6.3.2 Finding the Optimal Arrival Rate

When the arrival rate of the updates is a parameter that we can control, it is interesting to have an idea on its value that minimizes the average age.

Theorem 6.2. *For the M/G/1/1 blocking system, the minimum average age Δ^* is achieved for:*

- *If $C_S > 1$, then $\lambda^* = \frac{\beta^*}{(1-\beta^*)\mathbb{E}(S)}$ with $\beta^* = \sqrt{\frac{2}{C_S+1}}$ and*
$$\Delta^* = \mathbb{E}(S)\sqrt{2(C_S+1)}$$
- *If $C_S \leq 1$, $\lambda^* \to \infty$ and $\Delta^* = \mathbb{E}(S)\left(\frac{1}{2}(C_S+1)+1\right)$*

Proof. Setting the derivative of (6.4) with respect to β to zero, we get:
$$\beta^{*2} = \frac{2}{C_S+1}, \tag{6.7}$$
where β^* is the optimal value of β. Since $0 \leq \beta^* = \frac{\rho^*}{\rho^*+1} < 1$, C_S has to be strictly bigger than 1 for β^* to exist. In this case, $\beta^* = \sqrt{\frac{2}{C_S+1}}$ and solving for λ we get $\lambda^* = \frac{\beta^*}{(1-\beta^*)\mathbb{E}(S)}$. Using β^* in (6.4) gives the value of the minimum age Δ^*.

If the service time distribution is such that $C_S \leq 1$, then $\frac{\partial \epsilon}{\partial \beta} = -\frac{1}{\beta^2} + \frac{C_S+1}{2} < 0$. However, $\frac{\partial \beta}{\partial \lambda} = \frac{\mathbb{E}(S)}{(\lambda\mathbb{E}(S)+1)^2} \geq 0$. Therefore, $\frac{\partial \epsilon}{\partial \lambda} = \frac{\partial \epsilon}{\partial \beta}\frac{\partial \beta}{\partial \lambda} < 0$. Thus the average age is a strictly decreasing function of the arrival rate and the minimal average age is obtained as $\lambda \to \infty$. □

6.4 M/G/1/1 with Blocking HARQ System

Now, we study the effect of different HARQ policies on the average age when considering an M/G/1/1 queue with blocking. We assume that the updates are sent through a symbol erasure channel with erasure rate ϵ. Moreover, two HARQ protocols are visited: the infinite incremental redundancy (IIR) and the fixed redundancy (FR).

6.4.1 Infinite Incremental Redundancy

In this policy, an update consists of k_s information symbols and is encoded using a rateless code. This means that the monitor needs to receive at least k_s symbols in order to decode the update. The transmission of an update finishes whenever k_s symbols are successfully transmitted. All updates arriving when the system is busy are discarded. Therefore, we define the service time S of an update as the number of channel uses needed for the monitor to receive k_s symbols. Hence, S is distributed as a negative binomial with k_s successes and success probability $1-\epsilon$.

Theorem 6.3. *The average age of the M/G/1/1 blocking IIR-HARQ system is:*

$$\Delta_{NIIR} = \frac{1}{\lambda} + \frac{k_s}{1-\epsilon} + \frac{\lambda k_s(k_s+\epsilon)}{2(1-\epsilon)(\lambda k_s + 1 - \epsilon)}. \tag{6.8}$$

Moreover, the minimum average age is achieved for $\lambda \to \infty$ and its value is given by,

$$\Delta^*_{NIIR} = \frac{3k_s + \epsilon}{2(1-\epsilon)} \tag{6.9}$$

Proof. Since we are using IIR policy then the service time S of each update is distributed as a negative binomial $(k_s, 1-\epsilon)$, $S \in \{k_s, k_s+1, \ldots\}$. In this case the mean and variance of S are given by:

$$\mathbb{E}(S) = \frac{k_s}{1-\epsilon}, \quad \mathrm{Var}(S) = \frac{k_s\epsilon}{(1-\epsilon)^2}. \tag{6.10}$$

Hence, we compute the quantities ρ, β and C_S present in (6.4):

$$\rho = \frac{\lambda k_s}{1-\epsilon}, \quad \beta = \frac{\rho}{\rho+1} = \frac{\lambda k_s}{\lambda k_s + 1 - \epsilon}, \quad C_S = \frac{\epsilon}{k_s}. \tag{6.11}$$

Using the above expression in (6.4) and performing some simplifications we get (6.8).

Moreover, since $\epsilon \leq 1$ and $k_s \geq 1$, $C_S = \frac{\epsilon}{k_s} \leq 1$. By Theorem 6.2, the optimum average age is achieved as $\lambda \to \infty$. Taking the limit on (6.8) gives (6.9). □

6.4.2 Fixed Redundancy

In this policy, we apply two levels of coding: a packet level and a physical level. Each update consists of k_p packets encoded using a rateless code. This means that the monitor needs to receive k_p decodable packets in order to decode the update. Moreover, each packet contains k_s information symbols and is encoded using a (n_s, k_s)-Maximum Distance Separable (MDS) code. Hence, a packet can be decoded if at least k_s symbols are not erased. Since the packets are being transmitted through a symbol erasure channel with erasure probability ϵ than the probability for the receiver to be unable to decode a packet is:

$$\begin{aligned} \epsilon_p &= \mathbb{P}(\text{less than } k_s \text{ symbols received}) \\ &= \sum_{i=0}^{k_s-1} \binom{n_s}{i} \epsilon^{n_s-i}(1-\epsilon)^i. \end{aligned} \tag{6.12}$$

Theorem 6.4. *The average age of the M/G/1/1 FR-HARQ blocking system is*

$$\Delta_{NFR} = \frac{1}{\lambda} + \frac{n_s k_p}{1-\epsilon_p} + \frac{\lambda n_s^2 k_p(k_p+\epsilon_p)}{2(1-\epsilon_p)(\lambda n_s k_p + 1 - \epsilon_p)}. \tag{6.13}$$

Moreover, the minimum average age is achieved as $\lambda \to \infty$ and its value is given by,

$$\Delta^*_{NFR} = \frac{3n_s k_p + \epsilon_p}{2(1-\epsilon_p)} \tag{6.14}$$

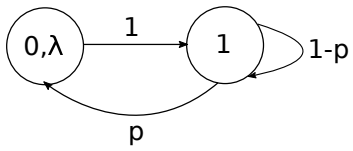


Figure 6.2 – Semi-Markov chain representing the queue for LCFS with preemption

Proof. The number M of packets needed to be transmitted to decode an update is distributed as a negative binomial $(k_p, 1-\epsilon_p)$ random variable with k_p successes and success rate $(1-\epsilon_p)$, $M \in \{k_p, k_p+1, \ldots\}$. Since the transmission of each packet consumes n_s channel uses then the service time S of each update is $S = n_s M$. Thus, the mean and variance of S are given by:

$$\mathbb{E}(S) = \mathbb{E}(n_s M) = n_s \mathbb{E}(M) = \frac{n_s k_p}{1-\epsilon_p}, \tag{6.15}$$

$$\mathrm{Var}(S) = \mathrm{Var}(n_s M) = n_s^2 \mathrm{Var}(M) = \frac{n_s^2 k_p \epsilon_p}{(1-\epsilon_p)^2}. \tag{6.16}$$

Hence, we compute the quantities:

$$\rho = \frac{\lambda k_p}{1-\epsilon_p}, \quad \beta = \frac{\lambda k_p}{\lambda k_p + 1 - \epsilon_p}, \quad C_S = \frac{\epsilon_p}{k_p}. \tag{6.17}$$

Using the above expressions in (6.4) and performing some simplifications we get (6.13).

Moreover, since $\epsilon_p \leq 1$ and $k_p \geq 1$, $C_S = \frac{\epsilon_p}{k_p} \leq 1$. By Theorem 6.2, the optimum average age is achieved as $\lambda \to \infty$. From (6.13) this yields (6.14). □

6.5 M/G/1/1 with Preemption

In the M/G/1/1 with preemption scenario, any packet being served is preempted if a new packet arrives and the new packet is served. In fact, while in the M/G/1/1 with blocking the priority is given to the update being served, in this setup, the priority goes to the newly generated update. Moreover, the number of packets in the queue can be modeled as a continuous-time two-state semi-Markov chain depicted in Figure 6.2. An interpretation of this chain can be found in Section 3.3, which we repeat here for convenience. The 0-state corresponds to empty queue and no packet is being served while the 1-state corresponds to the state where the queue is full and is serving one packet. However, given that the interarrival time between packets is exponentially distributed with rate λ then one spends an exponential amount of time X in the 0-state before jumping with probability 1 to the other state. Once in the 1-state, two independent clocks are started: the service time clock of the packet being served and the rate λ memoryless clock of the interarrival time between the current packet and the next one to be generated. If the memoryless clock ticks first, we stay in the 1-state, otherwise we go back to the 0-state. Hence, the jump from the 1-state to the 0-state occurs with probability $p = \mathbb{P}(S < X)$, where S is a generic service time with distribution $f_S(t)$ and X is a generic rate λ memoryless interarrival time which is independent of S.

Theorem 6.5. *The average age of an M/G/1/1 system with preemption with a single source is given by,*

$$\Delta = \frac{1}{\lambda P_\lambda}, \tag{6.18}$$

where $P_\lambda = \int_0^\infty f_S(t)e^{-\lambda t}\mathrm{d}t$ *is the Laplace transform of the service-time distribution.*

Proof. We obtain (6.18) by applying Theorem 4.1 and taking the number of sources $M = 1$. In Section 6.9, we present an alternative proof technique for this theorem. □

For the M/G/1/1 with preemption, the average age depends on the Laplace transform of the service time distribution.

6.6 M/G/1/1 with Preemption and HARQ

In this Section we study the effect of different HARQ policies on the average age when considering an M/G/1/1 queue with preemption. Indeed, we assume that the updates are sent through a symbol erasure channel with erasure rate ϵ. Moreover, two HARQ models are visited: the infinite incremental redundancy (IIR) and the fixed redundancy (FR).

6.6.1 Infinite Incremental Redundancy

In this setup, the transmission of an update finishes whenever one of these events happen first: (i) k_s symbols are successfully transmitted, or (ii) a new update is generated. Hence the following theorem.

Theorem 6.6. *The average age of an M/G/1/1 with preemption system when using the IIR policy is given by,*

$$\Delta_{PIIR} = \frac{1}{\lambda}\left(\frac{e^\lambda - \epsilon}{1-\epsilon}\right)^{k_s}. \tag{6.19}$$

Moreover, ϵ_{PIIR} *has a minimum and the arrival rate* λ^* *that achieves it should satisfy the condition*

$$\lambda^* \leq \frac{1}{k_s}. \tag{6.20}$$

The minimum age Δ^*_{PIIR} *can be lower bounded using*

$$\Delta^*_{PIIR} \geq \frac{1}{\lambda_{IIR}}\left(1 + \frac{\lambda_{IIR}}{1-\epsilon}\right)^{k_s}, \tag{6.21}$$

where $\lambda^* \approx \lambda_{IIR} = \frac{1-k_s+\sqrt{(k_s+1)^2-4k_s\epsilon}}{2k_s}$.

Proof. Under the IIR policy, the service time S of each update is distributed as a negative binomial $(k_s, 1-\epsilon)$, $S \in \{k_s, k_s+1, \ldots\}$. In this case the moment generating function of S is given by:

$$\phi_S(s) = \mathbb{E}\left(e^{sS}\right) = \left(\frac{1-e^s\epsilon}{e^s(1-\epsilon)}\right)^{-k_s}. \tag{6.22}$$

Noting that $P_\lambda = \phi_S(-\lambda)$ and using (6.18) and (6.22), we get (6.19). To prove condition (6.20) we differentiate Δ_{PIIR} with respect to λ and equate it to zero. This yields

$$-\frac{1}{\lambda}\left(\frac{e^\lambda-\epsilon}{1-\epsilon}\right)+\frac{k_s e^\lambda}{1-\epsilon}=0. \tag{6.23}$$

Thus, to satisfy (6.23) we need

$$e^\lambda(k_s\lambda-1)=-\epsilon. \tag{6.24}$$

Since $0\le\epsilon\le 1$, (6.24) implies that $k_s\lambda-1\le 0$. Hence (6.20) holds. Moreover, since $\lambda>0$, we have that $e^\lambda>1+\lambda$. This means that if λ^* minimizes Δ_{PIIR}, then

$$\Delta^*_{\text{PIIR}}=\Delta_{\text{PIIR}}(\lambda^*)>\frac{1}{\lambda^*}\left(1+\frac{\lambda^*}{1-\epsilon}\right)^{k_s}. \tag{6.25}$$

Finally, in order to obtain λ^* one needs to solve equation (6.24) which does not have a simple closed form expression. As an alternative, we can make the small λ approximation $e^{\lambda^*}\approx 1+\lambda^*$. In this case, (6.24) reduces to

$$(1+\lambda)(k_s\lambda-1)=-\epsilon. \tag{6.26}$$

This is a quadratic equation whose only positive root is given by

$$\lambda_{IIR}=\frac{1-k_s+\sqrt{(k_s+1)^2-4k_s\epsilon}}{2k_s}.$$

To obtain (6.21), we replace λ^* by λ_{IIR} in (6.25). □

Since $\lambda^*\le\frac{1}{k_s}\le 1$, the lower bound in (6.21) becomes a tight approximation of the average age for typical values of k_s.

6.6.2 Fixed Redundancy

In this case also the transmission of an update is terminated whenever one of these events happen first: (i) k_p packets are successfully transmitted, or (ii) a new update is generated. As in the M/G/1/1 blocking system, we define the packet erasure probability $\epsilon_p=\sum_{i=0}^{k_s-1}\binom{n_s}{i}\epsilon^{n_s-i}(1-\epsilon)^i$.

Theorem 6.7. *The average age of the information for an M/G/1/1 with preemption system using the FR policy is given by,*

$$\Delta_{PFR}=\frac{1}{\lambda}\left(\frac{1-e^{-\lambda n_s}\epsilon_p}{e^{-\lambda n_s}(1-\epsilon_p)}\right)^{k_p}. \tag{6.27}$$

Moreover, Δ_{PFR} has a minimum and the arrival rate λ^* that achieves it should satisfy the condition

$$\lambda^*\le\frac{1}{n_s k_p}. \tag{6.28}$$

The minimum age Δ^*_{PIIR} can be lower bounded using

$$\Delta^*_{PFR} \geq \frac{1}{\lambda_{FR}} \left(1 + \frac{\lambda_{FR} n_s}{1-\epsilon_p}\right)^{k_p}, \tag{6.29}$$

where $\lambda^* \approx \lambda_{FR} = \frac{1-k_p+\sqrt{(k_p+1)^2-4k_p\epsilon_p}}{2n_s k_p}$.

Proof. The number M of packets needed to be transmitted to decode an update is distributed as a negative binomial $(k_p, 1-\epsilon_p)$ random variable with k_p successes and success rate $(1-\epsilon_p)$, $M \in \{k_p, k_p+1, \ldots\}$. Since the transmission of each packet consumes n_s channel uses, the service time S of each update is $S = n_s M$. Thus, the moment generating function of S is:

$$\phi_S(s) = \mathbb{E}\left(e^{s n_s M}\right) = \phi_M(n_s s) = \left(\frac{e^{n_s s}(1-\epsilon_p)}{1-e^{n_s s}\epsilon_p}\right)^{k_p}. \tag{6.30}$$

Using (6.18), the fact that $P_\lambda = \phi_S(-\lambda)$ and the above expression we obtain (6.27).

To prove condition (6.28) we differentiate Δ_{PFR} with respect to λ and equate it to zero, yielding

$$-\frac{1}{\lambda}\left(\frac{e^{\lambda n_s}-\epsilon_p}{1-\epsilon_p}\right) + \frac{k_p n_s e^{\lambda n_s}}{1-\epsilon_p} = 0. \tag{6.31}$$

Thus, to satisfy (6.31) we need

$$e^{\lambda n_s}(k_p n_s \lambda - 1) = -\epsilon_p. \tag{6.32}$$

Since $0 \leq \epsilon_p \leq 1$, (6.32) implies that $k_p n_s \lambda - 1 \leq 0$. Hence (6.28) holds.

As in the proof for Theorem 6.6, here also we have:

$$\Delta^*_{\mathrm{PFR}} = \Delta_{\mathrm{PFR}}(\lambda^*) > \frac{1}{\lambda^*}\left(1 + \frac{n_s\lambda^*}{1-\epsilon_p}\right)^{k_p}. \tag{6.33}$$

Finally, also as in the proof for Theorem 6.6, we approximate the real value of λ^* by solving the quadratic equation

$$(1+\lambda n_s)(k_p n_s \lambda - 1) = -\epsilon_p. \tag{6.34}$$

The only positive root is given by

$$\lambda_{\mathrm{FR}} = \frac{1-k_p+\sqrt{(k_p+1)^2-4k_p\epsilon_p}}{2n_s k_p}.$$

To obtain (6.29), we replace λ^* by λ_{FR} in (6.33). □

Since $n_s\lambda^* \leq \frac{1}{k_p} \leq 1$, the lower bound in (6.29) becomes a tight approximation for typical values of k_p.

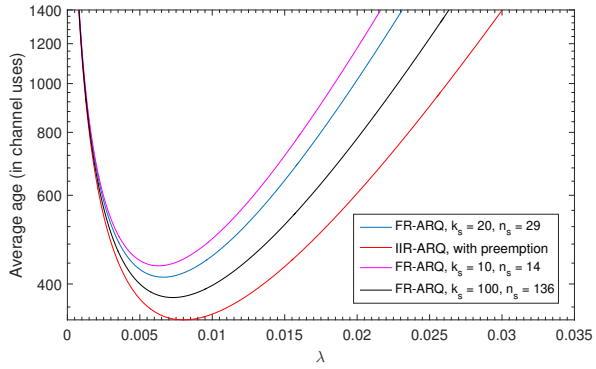


Figure 6.3 – Comparing the performance of the FR-HARQ for the M/G/1/1 with preemption scheme when varying the number of information symbols in each packet. We assume the update has 100 information symbols, $\epsilon = 0.2$, $k_p = 100/k_s$. n_s is chosen to minimize the average age.

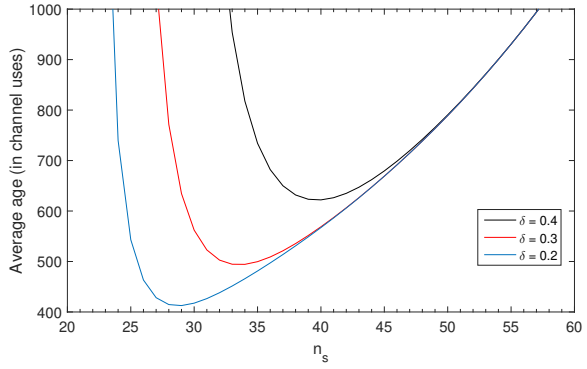


Figure 6.4 – Average age with respect to codeword length for the M/G/1/1 with preemption scheme with FR-HARQ. We assume the update has 100 information symbols, $\lambda = 0.0066$, $k_s = 20$ and $k_p = 100/k_s$.

6.7 Numerical Results

In this section, we first compare the two HARQ policies, IIR and FR, for the M/G/1/1 with and without preemption. Then, for each HARQ policy, we compare the performances of the two M/G/1/1 schemes. Moreover, for the simulation results discussed in this section, we assume the following setting: a symbol erasure channel with erasure rate $\epsilon = 0.2$ and each update in IIR-HARQ and FR-HARQ contain $K = 100$ information symbols. So for IIR-HARQ we have $f_s = 100$ while for FR-HARQ, we assume each update is divided into $k_p = K/k_s$ packets where each packet is encoded using an MDS-(k_s, n_s) code.

We first start analyzing the M/G/1/1 system with preemption. Fig. 6.3 shows the average age for different values of k_s around its minimum point. As we can notice, if we choose the optimum n_s for a fixed k_s and range of λ then the average age decreases as the number of packets per update decreases. In fact, the black curve which corresponds to $k_p = 1$ has the lowest average age around its minimum, followed

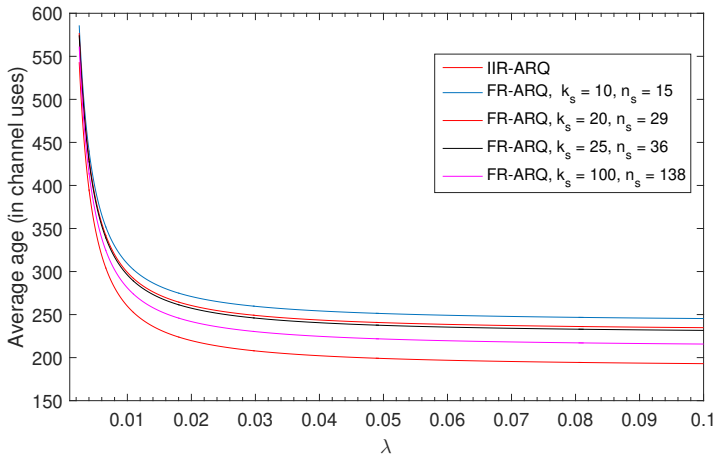


Figure 6.5 – Comparing the performance of the FR-HARQ for the M/G/1/1 without preemption scheme when varying the number of information symbols in each packet. We assume the update has 100 information symbols, $\epsilon = 0.2$, $k_p = 100/k_s$. n_s is chosen to minimize the average age.

by the blue curve associated with $k_p = 5$ and the worst performance is for the system with $k_p = 10$. Fig. 6.3 also confirms the results in Theorem 6.6 and Theorem 6.7 saying that ϵ_{PIIR} and ϵ_{PFR} achieve a minimum at a small value of λ. This figure also suggests that no matter how we choose k_s and n_s, IIR outperforms FR. The values of n_s chosen in Fig. 6.3 are such that they minimize the average age for a given ϵ and k_s. The existence of such optimum packet length in FR can be deduced from Fig. 6.4. Here we set $\lambda = 0.0066$, which minimizes the average age for $\epsilon = 0.2$, and $k_s = 20$. Fig. 6.4 can be explained using the lower bound (6.29): for a given λ, as n_s gets large, $\epsilon_p \to 0$ and the lower bound will be increasing with n_s since $\left(1 + \frac{n_s \lambda^*}{1-\epsilon_p}\right) > 1$. However, for n_s close to k_s, $\epsilon_p \to 1$ which also increases this lower bound. Thus, the packet length should be neither too small (equal to k_s) nor too large. As it is expected, Fig. 6.4 also shows that the optimal packet length n_s increases as the erasure rate ϵ increases.

The above results concerning the M/G/1/1 system with preemption apply also for the M/G/1/1 blocking system as it can be seen in Fig. 6.5 and 6.6. However, some differences need to be noted. (*i*) Fig. 6.5 confirms the results of Theorems 6.3 and 6.4 that the average age is a decreasing function of λ. (*ii*) Fig. 6.5 shows that for any value of λ, increasing the number of packets per update increases the average age. (*iii*) Fig. 6.6 shows the existence of an optimal packet length n_s for a given ϵ, λ and k_s.

Finally, we compare the performance of the M/G/1/1 with preemption and the M/G/1/1 blocking systems for each one of the HARQ policies. In both cases, Fig. 6.7 shows that the M/G/1/1 blocking system performs better than its counterpart for all values of λ.

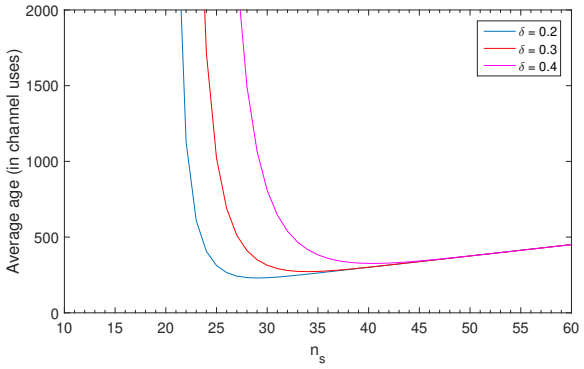


Figure 6.6 – Average age with respect to codeword length for the M/G/1/1 without preemption scheme with FR-HARQ. We assume the update has 100 information symbols, $\lambda = 1$, $k_s = 20$ and $k_p = 100/k_s$.

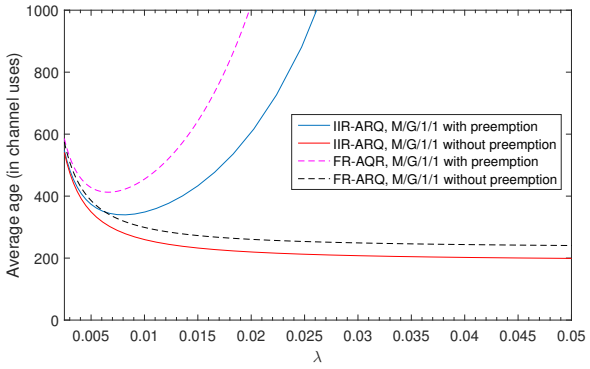


Figure 6.7 – Comparing the performance of the two M/G/1/1 schemes when using IIR and FR. We assume the update has 100 information symbols and $\epsilon = 0.2$.

6.8 Conclusion

In this chapter, we have studied the M/G/1/1 system along with the possible update management policies it presents: preempting the current update or discarding the newly generated one. We have derived general expressions for their average age and have used this result to compute the average age when considering a practical scenario: updates are sent over a symbol erasure channel using two different HARQ protocols, IIR and FR. In both cases, prioritizing the current update being sent and not preempting it turned out to be the best strategy. Moreover, as it is expected, the IIR protocol gives a better performance from an age point of view than FR. Finally, we have argued through simulations that for the FR protocol, ensuring reliable delivery of every update packet (by using large codeword length n_s) does not achieve the optimal average age.

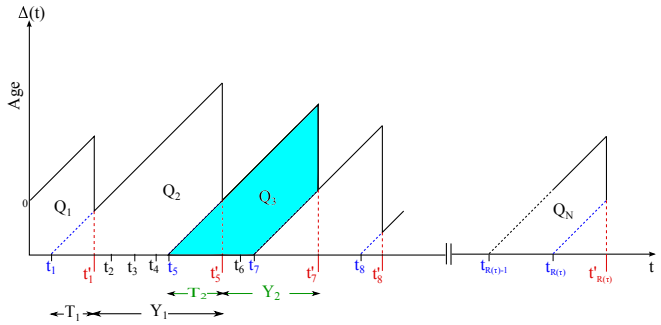


Figure 6.8 – Variation of the instantaneous age for LCFS with preemption

6.9 Appendix: Alternate Proof of Theorem 6.5

In this appendix we present an alternate method for computing the average age and average peak age of an M/G/1/1 with-preemption scheme. Whereas in Chapter 4 we computed these two age metrics for multiple sources, in this section we consider the system to have a single source.

The quantity $p = \mathbb{P}(S < X)$ will play an important role in our derivation, so we will take a closer look at it:

$$p = \int_0^\infty f_S(t)\mathbb{P}(X > t)\ \mathrm{d}t = \int_0^\infty f_S(t)e^{-\lambda t}\ \mathrm{d}t = P_\lambda, \tag{6.35}$$

where P_λ is the Laplace transform of the service time distribution.

Given an M/G/1/1 with-preemption scheme, Lemma 3.4 applies for any service-time distribution. Thus, we can apply it here also. This means that, for any i, Y_i and T_i are independent. Using Fig. 6.8 and (2.20) we have that the average age Δ is:

$$\Delta = \lambda_e \mathbb{E}(Q) = \lambda_e \left(\frac{1}{2}\mathbb{E}\left(Y^2\right) + \mathbb{E}(T)\mathbb{E}(Y) \right), \tag{6.36}$$

where T and Y as defined in Section 6.2. Using Lemma 3.1 and a similar proof to Lemma 3.3 we can write $\lambda_e = \lambda P_\lambda$ as the effective arrival rate. We start with $\mathbb{E}(T)$.

Lemma 6.2. *The PDF of the system time, T, of a successful update is*

$$f_T(t) = \frac{f_S(t)}{P_\lambda} e^{-\lambda t}. \tag{6.37}$$

Its expected value is

$$\mathbb{E}(T) = -\frac{1}{P_\lambda}\frac{\partial P_\lambda}{\partial \lambda}. \tag{6.38}$$

Proof.

$$\begin{aligned} f_T(t) &= \lim_{\epsilon \to 0} \frac{\mathbb{P}(S \in [t, t+\epsilon] | S < X)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\mathbb{P}(S \in [t, t+\epsilon])}{\epsilon P_\lambda} \mathbb{P}(S < X | S \in (t, t+\epsilon)) \\ &= \frac{f_S(t)}{P_\lambda}\mathbb{P}(X > t) = \frac{f_S(t)}{P_\lambda} e^{-\lambda t}. \end{aligned} \tag{6.39}$$

Using (6.37) we calculate the expected value of T:

$$\mathbb{E}(T) = \frac{1}{P_\lambda}\int_0^\infty t f_S(t) e^{-\lambda t} \mathrm{d}t = -\frac{1}{P_\lambda}\frac{\partial P_\lambda}{\partial \lambda}. \tag{6.40}$$

□

Now we only need to calculate the first and second moments of Y. For that we will derive its moment generating function.

Lemma 6.3. *The moment generating function of the interdeparture time Y is given by*

$$\phi_Y(s) = \frac{\lambda P_{\lambda-s}}{\lambda P_{\lambda-s} - s}, \tag{6.41}$$

where $P_{\lambda-s} = \int_0^\infty f_S(t) e^{-(\lambda-s)}\ \mathrm{d}t$.

Proof. From Fig. 6.8 we can deduce that Y is the shortest time to go from the 0-state back to the 0-state. This means that

$$Y = X + W, \tag{6.42}$$

where X is exponentially distributed with rate λ and W is

$$\begin{aligned} W &= \begin{cases} T & \text{with probability } p \\ X_1' + T & \text{with probability } (1-p)p \\ X_1' + X_2' + T & \text{with probability } (1-p)^2 p \\ \vdots \end{cases} \\ &= \sum_{j=0}^{M} X_j' + T, \end{aligned} \tag{6.43}$$

where $X_0' = 0$ and for $j > 0$, X_j' is such that $\mathbb{P}(X_j' < \alpha) = \mathbb{P}(X < \alpha | X < S)$. M, which gives the number of discarded packets before the first successful reception, is a geometric(p) random variable independent of X_j' and T. We start first by deriving the moment generating function of X'.

$$\begin{aligned} f_{X'}(t) &= \lim_{\epsilon \to 0} \frac{\mathbb{P}(X \in [t, t+\epsilon] | S > X)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\mathbb{P}X \in [t, t+\epsilon])}{\epsilon(1-P_\lambda)} \mathbb{P}(S > X | X \in (t, t+\epsilon)) \\ &= \frac{f_X(t)}{1-P_\lambda}\mathbb{P}(S > t) \\ f_{X'}(t) &= [1 - F_S(t)]\,\frac{\lambda e^{-\lambda t}}{1-P_\lambda}, \end{aligned} \tag{6.44}$$

where $F_S(t)$ is the cdf of the service time S. Hence,

$$\begin{aligned}\phi_{X'}(s) = \mathbb{E}\left(e^{sX'}\right) &= \int_0^\infty e^{st}\left(1 - F_S(t)\right)\frac{\lambda e^{-\lambda t}}{1-P_\lambda}\,\mathrm{d}t \\ &\overset{(a)}{=} \frac{\lambda}{\lambda - s}\frac{1}{1-P_\lambda} - \frac{\lambda}{1-P_\lambda}\frac{P_{\lambda-s}}{\lambda-s} \\ &= \frac{\lambda(1-P_{\lambda-s})}{(\lambda-s)(1-P_\lambda)},\end{aligned} \tag{6.45}$$

where (a) is obtained by using integration by parts with $u = 1 - F_S(t)$ and $\frac{\mathrm{d}v}{\mathrm{d}t} = e^{-t(\lambda-s)}$. On the other hand, (6.37) implies

$$\phi_T(s) = \mathbb{E}\left(e^{sT}\right) = \int_0^\infty \frac{f_S(t)}{P_\lambda}e^{-\lambda t}e^{st}\,\mathrm{d}t = \frac{P_{\lambda-s}}{P_\lambda}. \tag{6.46}$$

Using (6.45) and (6.46), we deduce the moment generating of W,

$$\begin{aligned}\phi_W(s) &= \mathbb{E}\left(e^{s\left(\sum_{i=0}^{M} X_i' + T\right)}\right) \\ &= \mathbb{E}\left(e^{sT}\right)\mathbb{E}\left(\mathbb{E}\left(e^{sX'}\right)^M\right) \\ &= \frac{P_{\lambda-s}}{P_\lambda}\sum_{i=0}^{\infty}\left(\frac{\lambda(1-P_{\lambda-s})}{(\lambda-s)(1-P_\lambda)}\right)^i (1-P_\lambda)^i P_\lambda \\ &= \frac{(\lambda-s)P_{\lambda-s}}{\lambda P_{\lambda-s} - s}.\end{aligned} \tag{6.47}$$

Using (6.47) and that $\phi_X = \mathbb{E}\left(e^{sX}\right) = \frac{\lambda}{\lambda-s}$, we get (6.41) from $\phi_Y(s) = \mathbb{E}\left(e^{sX}\right)\mathbb{E}\left(e^{sW}\right)$. □

Theorem 6.8. *The average age of an M/G/1/1 system with preemption is given by,*

$$\Delta = \lambda_e \mathbb{E}(Q) = \frac{1}{\lambda P_\lambda}. \tag{6.48}$$

Proof. Deriving (6.41) once and twice and setting $s = 0$ gives:

$$\mathbb{E}(Y) = \frac{1}{\lambda P_\lambda} \quad \text{and} \quad \mathbb{E}(Y^2) = \frac{2}{\lambda^2 P_\lambda^2}\left(1 + \lambda\frac{\partial P_\lambda}{\partial \lambda}\right) \tag{6.49}$$

Using (6.38) and (6.49) we get $\mathbb{E}(Q) = \frac{1}{\lambda^2 P_\lambda^2}$. This last expression and the fact that $\lambda_e = \lambda P_\lambda$ give (6.48). □

Optimal Age over Erasure Channels

7

7.1 Introduction

In the previous chapters we have mostly focused on computing the average age (AoI) and/or the average peak age (PAoI) given a certain status updating policy. In this chapter, we take an information-theoretic approach to the age problem and provide a characterization of the optimal achievable age when the channel used is the erasure channel and no feedback is assumed. This means that we consider the following question: Given an erasure channel with no feedback and with input alphabet $\mathcal{V}$ and a source with alphabet $\mathcal{U}$, what is the lowest average age that can be achieved in this system? To answer this problem we distinguish two cases:

- **Case 1**: The source alphabet and the channel-input alphabet are the same or there exists a bijection from $\mathcal{U}$ onto $\mathcal{V}$.
- **Case 2**: The source alphabet and the channel-input alphabet are different and of different size. This means that there is no bijection such that $\mathcal{V}$ is the image of $\mathcal{U}$ by this bijection.

For the first case we derive an exact closed-form expression for the average age and show that the optimal average age is achieved without any encoding done on the source symbols. Whereas for the second case, encoding is mandatory and we use random coding to give an upper and lower bounds on the achievable average age of the system, as well as an approximation of the lower bound inspired by [76, 77].

The rest of this chapter is organized as follows: In Section 7.2, we present the system model and some definitions which are common to all later sections. In Section 7.3, we derive the optimal average age for Case 1 and in Section 7.4 we characterize the optimal achievable region of the average age for Case 2.

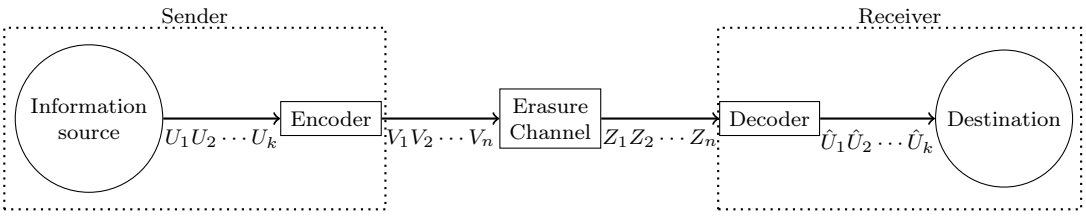


Figure 7.1 – The communication system.

7.2 Preliminaries

We first start by defining the communication system we will study. Fig. 7.1 illustrates such a system. The following discussion is based on the definitions presented in Section 1.1:

- *The channel*: We assume a discrete memoryless erasure channel with erasure probability ϵ. We refer to such channel by EC(ϵ). The channel-input alphabet is given by $\mathcal{V} = \{0, 1, \cdots, q-1\}$, q being a power of prime, and the channel-output alphabet by $\mathcal{V} \cup \{?\} = \{0, 1, \cdots, q-1, ?\}$. We also assume there is no feedback from the receiver. This means that the output of the encoder depends only on the source messages (or symbols) and the sender does not know whether a sent symbol was successfully received or not. In addition to that, we assume that transmitted channel-symbols are received instantaneously. However, there exists a period T_c between two consecutive channel uses. Thus, we define the *channel-use rate* $\mu = \frac{1}{T_c}$ to be the allowed number of channel uses per second.

- *The source*: We assume a single discrete memoryless source generating messages that belong to the set $\mathcal{U} = \{1, 2, \cdots, L\}$. So each symbol in this set is a message and we will use interchangeably the terms *source symbol* and *message* in this chapter. We also pose $k = \lceil \log_q(L) \rceil = \lceil \frac{\ln(L)}{\ln(q)} \rceil$ where $\log_q(x)$ for some $x > 0$ is the base-q logarithm of x. Hence, in order to represent one source symbol we need k channel-input symbols. This means that there exists an injective function $h(.)$ that maps every message $m \in \mathcal{U}$ to a length-k sequence $u^k \in \mathcal{V}^k$, with $u_j \in \mathcal{V}$ for $1 \leq j \leq k$. Thus, $h(\mathcal{U}) \subseteq \mathcal{V}^k$. Similar to the channel-use case, the source symbol generation is assumed periodic with period T_s. Thus, we define the *message generation rate* $\lambda = \frac{1}{T_s}$ as the fixed number of source symbols generated per second. Finally, the reader can notice that, when both the source alphabet and the channel-input alphabet have the same size, $h(\mathcal{U}) = \mathcal{V}$ and $k = 1$. In this case, we take $\mathcal{U} = \mathcal{V}$. In the case where the source alphabet and channel-input alphabet have different sizes, we focus on strategies induced by linear codes. For such strategies, without loss of generality and for ease of notation, we will consider the source alphabet $\mathcal{U} = \mathcal{V}^k$ and every message or source symbol to be a random sequence $U_1, U_2, \cdots, U_k$ chosen in an i.i.d fashion from $\mathcal{V}^k$.

- *The encoder and decoder*: At the i^{th} channel use, the encoder uses all the generated source symbols and encodes them into a single channel-input letter, $f_i : \mathcal{U}^{\lfloor \frac{iT_c}{T_s} \rfloor} \to \mathcal{V}$. The decoder, at the i^{th} channel use, uses all received channel-output

letters to output an estimate of the latest message that was fully transmitted, along with its index. Thus, $g_i : (\mathcal{V} \cup \{?\})^i \to (\mathcal{U} \cup \{\text{erasure}\}) \times \{1, 2, \ldots, \lfloor \frac{iT_c}{T_s} \rfloor\}$.

In the previous section, we indicated that we are interested in bounding the optimal achievable average age. Here, we define the concepts of achievable age and optimal achievable age.

Definition 7.1. *We call $\mathcal{C} = (f_i, g_i)_{i\geq 1}$ to be a coding scheme where $(f_i)_{i\geq 1}$ is the sequence of encoders and $(g_i)_{i\geq 1}$ is the sequence of decoders. The average age corresponding to such scheme is denoted by*

$$\Delta_{\mathcal{C}} = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau \Delta_{\mathcal{C}}(t)\mathrm{d}t,$$

where $\Delta_{\mathcal{C}}(t)$ is the instantaneous age.
Such a definition is independent of the choice of the channel. However, for the special case of the erasure channel with erasure probability ϵ, the average age relative to the coding scheme $\mathcal{C}$ will be denoted by

$$\Delta_{\epsilon,\mathcal{C}} = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau \Delta_{\epsilon,\mathcal{C}}(t)\mathrm{d}t, \tag{7.1}$$

where $\Delta_{\epsilon,\mathcal{C}}(t)$ is the instantaneous age.

Definition 7.2. *We say that an age D is achievable for an erasure channel with erasure probability ϵ (EC(ϵ)), if $\forall\delta > 0$ there exists a coding scheme $\mathcal{C} = (f_i, g_i)_{i\geq 1}$[1] such that*

$$\Delta_{\epsilon,\mathcal{C}} \leq D + \delta, \tag{7.2}$$

and the error probability on the decoded messages is zero.

Definition 7.3. *Given a channel EC(ϵ), we define the optimal average age Δ_ϵ to be the minimum achievable average age. Formally,*

$$\Delta_\epsilon = \inf_{\mathcal{C}\in\Gamma} \Delta_{\epsilon,\mathcal{C}}, \tag{7.3}$$

where Γ is the set of all possible coding schemes.
The set $\mathcal{R} = \{(\epsilon, D); D \geq \Delta_\epsilon \text{ and } \epsilon \in [0, 1]\}$ forms the set of achievable average ages over all erasure channels.

7.3 Optimal Age with the Same Source & Channel Alphabets

In this first case, we take $k = 1$ which means that the source and channel-input alphabets are the same. We first show that to achieve the optimal age, no encoding is required and provide the optimal transmission policy. We then compute the optimal average age.

[1] Given that the source alphabet is usually fixed, the only variables left to tune in any coding scheme are the blocklength, the encoder and the decoder.

7.3.1 The Optimal Transmission Policy

Theorem 7.1. *For a channel EC(ϵ), if the source alphabet and the channel-input alphabet are the same, then in order to minimize the average age no encoding is required.*

Proof. If an oracle were to give the erasure pattern to the transmitter then the optimal thing to do from an age perspective is to send the newest source symbol at the non-erased channel-uses since each message needs only one channel use to be transmitted. This means that at every channel use the sender is sending the freshest information. If instead we encode the messages using a coding scheme $\mathcal{C}$ into codewords of length $n > 1$ then each message will need strictly more than one channel use to be transmitted, hence, at any instant t and for an arbitrary erasure pattern, the instantaneous age would be larger than the one that corresponds to the uncoded messages, i.e. $\Delta_{\epsilon,\mathcal{C}}(t) \geq \Delta_{\epsilon,uncoded}(t)$. Therefore, $\Delta_{\epsilon,\mathcal{C}} \geq \Delta_{\epsilon,uncoded}$. This is so since with encoded messages, the transmitter is not necessarily sending the freshest information at every channel use. □

Theorem 7.2. *For a channel EC(ϵ), if the source alphabet and the channel-input alphabet are the same, then the optimal stable transmission policy from an age perspective is to keep transmitting the last-generated source-symbol until a new one is generated and discard all previous messages. This is a LCFS system with no buffer policy.*

Proof. Let's assume that an oracle provide us with the erasure pattern. From Theorem 7.1 we know that we should not encode the source symbols. It is clear that at each non-erased channel use we should send the latest update so that the drop in the instantaneous age is the most important. Indeed, if there is a non-erased channel use at time t' and the latest update is generated at t_{last} then the instantaneous age, $\Delta_{opt}(t)$, that corresponds to the LCFS with no buffer policy drops to $\Delta_{opt}(t) = t' - t_{last}$. If, instead, we use a different policy Π' where we send at t' any message generated at time $t < t_{last}$ then its instantaneous age $\Delta_{\Pi'}(t) = t' - t > \Delta_{opt}(t)$. This means that from t' onward $\Delta_{\Pi'}(t)$ will be point-wise larger or equal to $\Delta_{opt}(t)$, hence the average age $\Delta_{\Pi'} \geq \Delta_{opt}$. This argument shows that the optimal transmission policy would send the latest generated source symbol at every non-erased channel use while it can transmit anything at the erased channel uses. However, since in practice the transmitter do not have access to the erasure pattern beforehand, the policy that consists of keeping on transmitting the last generated update until a new one is created satisfies the optimality criterion that is to send the latest generated message at each non-erased channel use.

For the case where $T_c \leq T_s$ or $\mu \geq \lambda$, the LCFS with no buffer policy leads to the transmission of all source symbols at least once. Whereas for the case of $T_c > T_s$ or $\mu < \lambda$, some messages will be dropped since they will never be sent. □

7.3.2 The Optimal Average Age

Theorem 7.3. *Given an erasure channel EC(ϵ) with channel-use rate μ, a source with message-generation rate λ and utilization $\rho = \frac{\lambda}{\mu}$, then the optimal average age achieved over EC(ϵ) is:*

- *For irrational utilization $\rho \in \mathbb{R} \setminus \mathbb{Q}$,*

$$\Delta_\epsilon = \frac{1}{2\lambda} + \frac{1+\epsilon}{2\mu(1-\epsilon)}. \tag{7.4}$$

- *For rational utilization $\rho \in \mathbb{Q}$,*

$$\Delta_\epsilon = \frac{l-1}{2l\lambda} + \frac{1+\epsilon}{2\mu(1-\epsilon)}, \tag{7.5}$$

  where $\rho = \frac{\lambda}{\mu}$ can be written in an irreducible form $\rho = \frac{m}{l}$ with $m, l \in \mathbb{N}$, $l \neq 0$ and $\gcd(m, l) = 1$.

Before giving the proof of Theorem 7.3, we need the following lemmas.

Lemma 7.1. *Let $\alpha \in \mathbb{R}$ be an irrational number. Then for $n \in \mathbb{N}$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} [i\alpha] = \frac{1}{2}, \tag{7.6}$$

where $[i\alpha] = i\alpha - \lfloor i\alpha \rfloor$ is the fractional part of $i\alpha$.

Proof. This lemma is a consequence of Weyl's equidistribution theorem [69]. A full proof of this lemma and an overview of the equidistribution theory behind it are presented in Section 7.6. □

Lemma 7.2. *Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be a positive rational number. Moreover, assume that α can be written in an irreducible form as $\alpha = \frac{m}{l}$, where $m, n \in \mathbb{N}$, $l \neq 0$ and $\gcd(m, l) = 1$. Then, for $n \in \mathbb{N}$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} [i\alpha] = \frac{l-1}{2l}, \tag{7.7}$$

where $[i\alpha] = i\alpha - \lfloor i\alpha \rfloor$ is the fractional part of $i\alpha$.

Proof. A full proof of this lemma is presented in Section 7.6.3. □

Proof of Theorem 7.3. In this proof we will use a different approach than the ATA and DTA presented in Chapter 2. We know that $\mu = \frac{1}{T_c}$. In Chapter 1, we saw that the average age is given by (1.2) which is

$$\Delta_\epsilon = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau \Delta_\epsilon(t) \mathrm{d}t.$$

Noticing that $\tau = \frac{\tau}{T_c} T_c = \left(\lfloor \frac{\tau}{T_c} \rfloor + \left[\frac{\tau}{T_c} \right] \right) T_c$, with $\left[\frac{\tau}{T_c} \right]$ being the fractional part of $\frac{\tau}{T_c}$, we can rewrite the average age as

$$\begin{aligned}
\Delta_\epsilon &= \lim_{\tau \to \infty} \frac{1}{\tau} \left(\sum_{n=1}^{\lfloor \frac{\tau}{T_c} \rfloor} \int_{(n-1)T_c}^{nT_c} \Delta_\epsilon(t) \mathrm{d}t + \int_{\lfloor \frac{\tau}{T_c} \rfloor T_c}^{\tau} \Delta_\epsilon(t) \mathrm{d}t \right) \\
&= \lim_{\tau \to \infty} \frac{1}{\tau} \sum_{n=1}^{\lfloor \frac{\tau}{T_c} \rfloor} \int_{(n-1)T_c}^{nT_c} \Delta_\epsilon(t) \mathrm{d}t \\
&= \lim_{\tau \to \infty} \frac{T_c}{\tau} \sum_{n=1}^{\lfloor \frac{\tau}{T_c} \rfloor} \frac{1}{T_c} \int_{(n-1)T_c}^{nT_c} \Delta_\epsilon(t) \mathrm{d}t. \qquad (7.8)
\end{aligned}$$

The second equality is due to the fact that the instantaneous age $\Delta_\epsilon(t)$ is bounded over a finite interval. This means that for $t \in [\lfloor \frac{\tau}{T_c} \rfloor T_c, \tau]$, there exists a positive real number $L > 0$ such that

$$\max_{t \in [\lfloor \frac{\tau}{T_c} \rfloor T_c, \tau]} \Delta_\epsilon(t) < L.$$

Given that $\tau - \lfloor \frac{\tau}{T_c} \rfloor T_c < T_c$, then

$$0 \leq \frac{1}{\tau} \int_{\lfloor \frac{\tau}{T_c} \rfloor T_c}^{\tau} \Delta_\epsilon(t) \mathrm{d}t \leq \frac{1}{\tau} L T_c.$$

Hence,

$$\lim_{\tau \to \infty} \frac{1}{\tau} \int_{\lfloor \frac{\tau}{T_c} \rfloor T_c}^{\tau} \Delta_\epsilon(t) \mathrm{d}t = 0.$$

Let

$$M_\tau = \lfloor \frac{\tau}{T_c} \rfloor \quad \text{and} \quad \Delta_{\epsilon,n} = \frac{1}{T_c} \int_{(n-1)T_c}^{nT_c} \Delta_\epsilon(t) \mathrm{d}t.$$

Therefore, because

$$\lim_{\tau \to \infty} \frac{M_\tau}{\frac{\tau}{T_c}} = 1,$$

(7.8) becomes

$$\Delta_\epsilon = \lim_{\tau \to \infty} \frac{1}{M_\tau} \sum_{n=1}^{M_\tau} \Delta_{\epsilon,n}. \qquad (7.9)$$

At time $t \in [(n-1)T_c, nT_c]$,

$$\Delta_\epsilon(t) = t - u(t) = t - \frac{1}{\lambda} \lfloor \frac{\lambda}{\mu} \left(\lfloor t\mu \rfloor - K_n \right) \rfloor,$$

where $u(t)$ is the timestamp of the last successfully received source symbol at time t and K_n is the number of transmissions since the newest reception instant before time nT_c of the last successfully received source symbol. For example, if $K_n = 0$, this

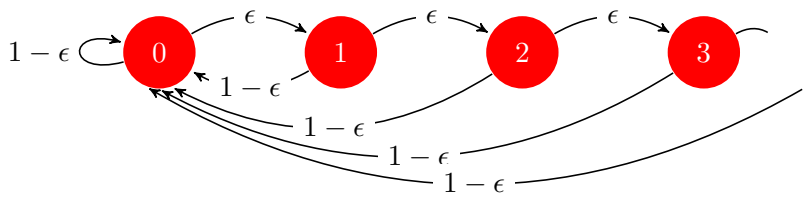


Figure 7.2 – Markov chain governing the number of transmissions since the reception instant of the last successful source symbol.

means that the last successful source symbol was received at the channel-use slot directly prior to t, in other words it was received at time $(n-1)T_c$. Hence,

$$\begin{aligned}
\Delta_{\epsilon,n} &= \frac{1}{T_c}\int_{(n-1)T_c}^{nT_c} \Delta_\epsilon(t)\mathrm{d}t \\
&= \frac{1}{T_c}\int_{(n-1)T_c}^{nT_c} \left(t - \frac{1}{\lambda}\lfloor\frac{\lambda}{\mu}\left(\lfloor t\mu\rfloor - K_n\right)\rfloor\right)\mathrm{d}t \\
&= \frac{1}{T_c}\left(\int_{(n-1)T_c}^{nT_c} t\mathrm{d}t - \frac{T_c}{\lambda}\lfloor\frac{\lambda}{\mu}\left(n-1-K_n\right)\rfloor\right) \\
&= \frac{1}{T_c}\left(nT_c^2 - \frac{T_c^2}{2} - \frac{T_c}{\lambda}\lfloor\frac{\lambda}{\mu}\left(n-1-K_n\right)\rfloor\right) \\
&= nT_c - \frac{T_c}{2} - \frac{1}{\lambda}\lfloor\frac{\lambda}{\mu}\left(n-1-K_n\right)\rfloor \\
&= \frac{1}{\lambda}\left(\frac{\lambda}{\mu}(n-1-K_n) - \lfloor\frac{\lambda}{\mu}\left(n-1-K_n\right)\rfloor\right) - \frac{1}{\mu}(n-1-K_n) + \frac{n}{\mu} - \frac{1}{2\mu} \\
&= \frac{1}{\lambda}\left(\frac{\lambda}{\mu}(n-1-K_n) - \lfloor\frac{\lambda}{\mu}\left(n-1-K_n\right)\rfloor\right) - \frac{1}{\mu}(n-1-K_n) + \frac{n}{\mu} - \frac{1}{2\mu} \\
&= \frac{1}{2\mu} + \frac{K_n}{\mu} + \frac{1}{\lambda}\left(\frac{\lambda}{\mu}(n-1-K_n) - \lfloor\frac{\lambda}{\mu}\left(n-1-K_n\right)\rfloor\right) \qquad (7.10)
\end{aligned}$$

where the third equality is due to the following fact: for all $t \in [(n-1)T_c, nT_c]$, $\lfloor t\mu\rfloor = n-1$.

Setting $K_1 = 0$, then for $n \geq 2$ we can write K_n as

$$K_n = \begin{cases} K_{n-1}+1 & \text{with probability } \epsilon \\ 0 & \text{with probability } 1-\epsilon. \end{cases}$$

So the K_n's form a Markov process represented by the Markov chain in Fig. 7.2. This Markov process is ergodic and has a stationary distribution which is identical to a geometric random variable K. This means,

$$\begin{aligned}
&\mathbb{P}(K_n = 0) = \mathbb{P}(K = 0) = 1-\epsilon \\
&\mathbb{P}(K_n = i) = \mathbb{P}(K = i) = \epsilon^i(1-\epsilon) \quad \forall\ i \geq 1
\end{aligned}$$

and

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} K_n = \mathbb{E}(K) = \frac{\epsilon}{1-\epsilon}. \qquad (7.11)$$

Replacing (7.10) in (7.9), we get

$$\begin{aligned}
\Delta_\epsilon &= \lim_{\tau\to\infty} \frac{1}{M_\tau} \sum_{n=1}^{M_\tau} \left(\frac{1}{2\mu} + \frac{K_n}{\mu} + \frac{1}{\lambda} \left(\frac{\lambda}{\mu}(n-1-K_n) - \lfloor \frac{\lambda}{\mu}(n-1-K_n) \rfloor \right) \right) \\
&= \lim_{\tau\to\infty} \frac{1}{M_\tau} \sum_{n=1}^{M_\tau} \left(\frac{1}{\lambda} \left(\frac{\lambda}{\mu}(n-1-K_n) - \lfloor \frac{\lambda}{\mu}(n-1-K_n) \rfloor \right) + \frac{1}{2\mu} + \frac{K_n}{\mu} \right) \\
&= \frac{1}{2\mu} + \frac{1}{\mu} \lim_{\tau\to\infty} \frac{1}{M_\tau} \sum_{n=1}^{M_\tau} K_n + \frac{1}{\lambda} \lim_{\tau\to\infty} \frac{1}{M_\tau} \sum_{n=1}^{M_\tau} \left(\frac{\lambda}{\mu}(n-1-K_n) - \lfloor \frac{\lambda}{\mu}(n-1-K_n) \rfloor \right) \\
&= \frac{1}{2\mu} + \frac{\mathbb{E}(K)}{\mu} + \frac{1}{\lambda} \lim_{\tau\to\infty} \frac{1}{M_\tau} \sum_{n=1}^{M_\tau} \left(\frac{\lambda}{\mu}(n-1-K_n) - \lfloor \frac{\lambda}{\mu}(n-1-K_n) \rfloor \right) \\
&= \frac{1}{2\mu} + \frac{\epsilon}{\mu(1-\epsilon)} + \frac{1}{\lambda} \lim_{\tau\to\infty} \frac{1}{M_\tau} \sum_{n=1}^{M_\tau} \left[\frac{\lambda}{\mu}(n-1-K_n) \right] \qquad (7.12)
\end{aligned}$$

where the last two equalities are justified by (7.11) and $\left[\frac{\lambda}{\mu}(n-1-K_n)\right]$ is the fractional part of $\frac{\lambda}{\mu}(n-1-K_n)$. Observe that $\forall\ n \geq 1$, $K_n \leq n-1$ and as $\tau \to \infty$, $M_\tau \to \infty$. Therefore, it is clear that the residual process

$$R_n = (n-1-K_n)_{n\geq 1}$$

and the process $(K_n)_{n\geq 1}$ are identically distributed [16]. In addition to that, given that $(K_n)_{n\geq 1}$ is ergodic then so is $(R_n)_{n\geq 1}$.

$$\begin{aligned}
\lim_{\tau\to\infty} \frac{1}{M_\tau} \sum_{n=1}^{M_\tau} \left[\frac{\lambda}{\mu}(n-1-K_n) \right] &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \left[R_n \frac{\lambda}{\mu} \right] \\
&= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \left[n \frac{\lambda}{\mu} \right].
\end{aligned}$$

At this point, we need to distinguish between two cases:

- $\rho = \frac{\lambda}{\mu}$ is irrational. Then, by Lemma 7.1,

  $$\lim_{M_\tau\to\infty} \frac{1}{M_\tau} \sum_{n=1}^{M_\tau} \left[n \frac{\lambda}{\mu} \right] = \frac{1}{2}.$$

  Using this result in (7.12) we get (7.4).

- $\rho = \frac{\lambda}{\mu} = \frac{m}{l}$ is rational with $m, l \in \mathbb{N}$, $l \neq 0$ and $\gcd(m,l) = 1$. Then, by Lemma 7.2,

  $$\lim_{M_\tau\to\infty} \frac{1}{M_\tau} \sum_{n=1}^{M_\tau} \left[n \frac{\lambda}{\mu} \right] = \frac{l-1}{2l}.$$

  Using this result in (7.12) we get (7.5).

□

7.4 Optimal Age with Different Source & Channel alphabets

In this setup, we consider the model described Section 7.2: The channel is a $q-ary$ erasure channel EC(ϵ) with no feedback and an input alphabet $\mathcal{V} = \{0, 1, \cdots, q-1\}$. The source alphabet is $\mathcal{U} = \mathcal{V}^k$ where $k > 1$. In addition to that we consider a special case where $\lambda = \mu$, so that at every channel use, a new source symbol is generated. Without loss of generality, let's assume $\lambda = \mu = 1$. The difference between the source alphabet and the channel-input alphabet as well as the presence of erasures impose the use of channel coding on the generated source symbols before their transmissions. However, the absence of feedback forces us to use fixed blocklength codes so that the transmitter and the receiver are always synchronized. We will focus on coding schemes that are induced by linear codes, so we will assume that $\mathcal{V} = \mathbb{F}_q$, where q is a power of a prime number. More precisely, for the l^{th} message to be transmitted, let $f_l : \mathcal{V}^k \to \mathcal{V}^n$ be an (n, k) linear code. We sequentially transmit messages generated from the source as follows: Assume that the j^{th} message to be transmitted (which belongs to $\mathcal{V}^k$) is generated at the instant t_l and discard all previous messages. We encode this source symbol using f_l, and obtain n symbols in $\mathcal{V}$. Finally, we transmit the n $\mathcal{V}$-symbols during the next n channel uses (i.e., at $t_l, \ldots, t_l + n - 1$). Fig. 7.1 illustrates this concept. The encoders $(f_l)_{l\geq 1}$ used to encode different messages can be different. This means that the encoder-decoder pairs $(f_l, g_l)_{l\geq 1}$ can be different. We denote a coding scheme defined by a given sequence of $(f_l, g_l)_{l\geq 1}$ that is induced by (n, k) linear codes as $\mathcal{C}(n, k)$.

7.4.1 The Optimal Transmission Policy

Definition 7.4. *An $(n, k)-$linear code is called* maximum distance separable *(MDS) if it achieves the Singleton bound:*

$$d = n - k + 1,$$

with d denoting the minimum distance[2] between the codewords of the code.

Proposition 7.1. *If the encoder f generates an MDS code, then*

- *any k columns of the generator matrix $\mathbf{G}$ are linearly independent,*
- *any subset of size k taken from a length-n codeword is sufficient to recover, with probability 1, the transmitted message.*

This means that if the channel is an EC(ϵ), the decoder needs to observe only k unerased channel-input symbols in order to perfectly decode the transmitted source symbol.

The proof of Proposition 7.1 is outside the scope of this text and we refer the reader to [55] for more details. The following theorem presents the optimal channel codes from an age point of view when the channel does not have any feedback.

[2] See [55] for more details.

Theorem 7.4. *Consider an EC(ϵ) with no feedback. Among all fixed-blocklength linear codes, MDS codes are age optimal. This means, to achieve age optimality, all codes used in the scheme $\mathcal{C}(n,k)$ should be MDS.*

Proof. Fix two positive integers k and n, an erasure pattern $\mathcal{E}$. Consider there exists a coding scheme $\mathcal{C}(n,k)$ where all encoders $(f_{\mathcal{C},l})_{l\geq 1}$ use MDS codes and another scheme $\mathcal{C}'(n,k)$ that uses the same encoders as $\mathcal{C}(n,k)$ except for the p^{th} message for which the encoder, $f_{\mathcal{C}',p}$, uses a non-MDS linear code. This means that $(f_{\mathcal{C},l})_{l\neq p} = (f_{\mathcal{C}',l})_{l\neq p}$. Notice that, for the p^{th} message encoded using $f_{\mathcal{C},p}$ and $f_{\mathcal{C}',p}$ and for fixed n and k, the MDS code has a minimum distance $d_{\mathcal{C},p} = n-k+1$ larger than that of any non-MDS linear code used by $f_{\mathcal{C}',p}$. This means that if the decoder g correctly decodes the channel output Z^n using $f_{\mathcal{C}',p}$, then it will certainly decode Z^n correctly using $f_{\mathcal{C},p}$. However, the converse is not true. Now consider a source with alphabet $\mathcal{V}^k$ generating messages and sending them through two parallel ECs(ϵ) but with the same erasure pattern $\mathcal{E}$. For the first channel we use the coding scheme $\mathcal{C}'(n,k)$, while for the second channel we use the coding scheme $\mathcal{C}(n,k)$. We assume the scenario where both schemes $\mathcal{C}(n,k)$ and $\mathcal{C}'(n,k)$ decode correctly exactly the same messages. This is a worst case scenario from the scheme $\mathcal{C}$ point of view since it can do better. However, even with such an assumption, the instantaneous age $\Delta_{\epsilon,\mathcal{C}}(t)$, relative to the scheme $\mathcal{C}$, is pointwise smaller or equal than the instantaneous age $\Delta_{\epsilon,\mathcal{C}'}(t)$ relative to the scheme $\mathcal{C}'$. We use Fig. 7.3 to prove this claim: Given that the transmission of each message takes exactly n channel uses and that the successful updates (source symbols) are the same whether using $\mathcal{C}$ or $\mathcal{C}'$, then if the transmission of the p^{th} update begins at time t_1, we have $\Delta_{\epsilon,\mathcal{C}}(t_1) = \Delta_{\epsilon,\mathcal{C}'}(t_1)$. If the p^{th} message is declared erased by both schemes, both instantaneous ages behave similarly and increase by $n\mu = n$ seconds. In the case of a successful transmission of the p^{th} message, $\Delta_{\epsilon,\mathcal{C}'}(t)$ increases linearly till the n^{th} channel use, at which instant it drops to $(n-1)\mu = n-1$ seconds. Whereas, $\Delta_{\epsilon,\mathcal{C}}(t)$ increases linearly till the k^{th} successful channel use, at which instant it drops to $(k-1)\mu = k-1$ seconds. $\Delta_{\epsilon,\mathcal{C}}(t)$ then increases linearly to $n-1$ seconds at the end of the transmission. Since $k \geq n$, this means that $\Delta_{\epsilon,\mathcal{C}}(t) \leq \Delta_{\epsilon,\mathcal{C}'}(t)$, $\forall\, t > 0$. This result also implies that the average age relative to $\mathcal{C}$, $\Delta_{\epsilon,\mathcal{C}}$, is smaller or equal to $\Delta_{\epsilon,\mathcal{C}'}$, the average age relative to $\mathcal{C}'$. □

Now we are ready to present the optimal transmission policy for a coding scheme $\mathcal{C}(n,k)$ with $\lambda = \mu = 1$.

Lemma 7.3. *For $\lambda = \mu$, at every channel use, a new source symbol is generated. The optimal transmission scheme from an age point of view is to use an MDS codes, and, whenever a message finishes its transmission after n channel uses, we begin transmitting, at the $(n+1)^{th}$ channel use, the source symbol generated at that same time instant. This means that when the packet (or message) generated at time t_0 finishes its transmission at time $t' = t_0 + n - 1$, the next packet to be transmitted at $t_0 + n$ is the one generated at this same time instant. All messages generated between $t_0 + 1$ and $t_0 + n - 1$ are dropped.*

Proof. Storing any number of packets and sending them will lead to a non-zero waiting time incurred by the stored messages since $n \geq k > 1$. Whereas the policy

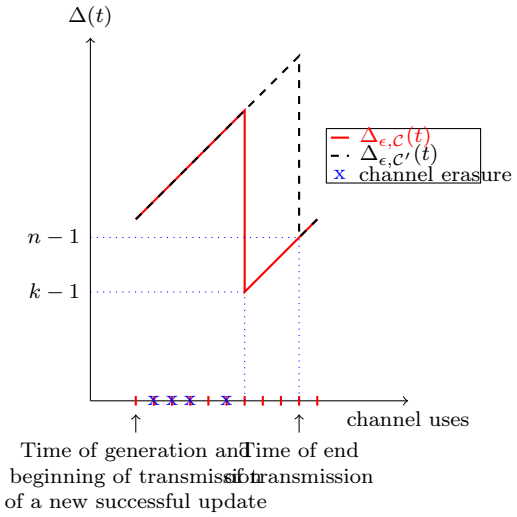


Figure 7.3 – Variation of the instantaneous age for an MDS code $\mathcal{C}$ and a non-MDS code $\mathcal{C}'$. We assume the erasure probability $\epsilon = 0.4$, $n = 10$ and $k = 3$.

presented in Lemma 7.3 guarantees zero waiting time and whenever the system is idle, the freshest message is transmitted. Moreover, since we are assuming a constant service-time, then using [67, Lemma 3] we deduce that the just-in-time policy is optimal. In our case, the just-in-time policy translates into the scheme described in Lemma 7.3. □

Theorem 7.4 shows that for a given couple (n, k), the optimal coding scheme is the one that uses only MDS codes. However, an explicit construction of such codes is not available for all values of (n, k). In the rest of this paper, we use random codes to give an upper bound on the optimal average age achieved using MDS codes. The use of random coding to construct fountain-like codes was used by Shamai et al. in [62]. In this paper, the authors show that without any randomness we cannot properly define the notion of fountain capacity because there is always a case where the deterministic fountain codes cannot achieve any positive rate with an error probability tending to 0. Nevertheless, we use the rateless (or fountain) codes, previously adopted in Chapter 6, to give a lower bound on the optimal achievable average-age Δ_ϵ. As shown in [62], these codes cannot be implemented in practice, that's why we do not consider them part of the possible coding schemes.

7.4.2 The Random Code

Before presenting our age analysis, we begin by defining the encoder and decoder used in this setup. Consider a $\mathcal{C}(n, k)$ coding scheme. The encoder-decoder pair (f_l, g_l), corresponding to the l^{th} message to be transmitted, is constructed as follows: Since we are interested in linear codes, we use the generator matrix in order to create our code. For that, we choose the n columns of the generator matrix $\mathbf{G}_l$ independently and uniformly at random from the set $\mathcal{V}^k \setminus \{0^k\}$, where 0^k is the sequence of k zeros.

We denote by $(\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_n)$ [3] the n columns of $\mathbf{G}_l$. Thus

$$\mathbf{G}_l = \begin{bmatrix} \mathbf{g}_1 & \mathbf{g}_2 & \cdots & \mathbf{g}_n. \end{bmatrix}$$

Once this matrix is generated, it is shared between the encoder and the decoder. For each new message to be transmitted, we generate a new generator matrix. However, the encoder and decoder work in a similar fashion for all messages:

- Let $\mathbf{u} = u^k = \begin{bmatrix} u_1 & u_2 & \cdots & u_k \end{bmatrix} \in \mathcal{V}^k$ be the l^{th} message to be sent. Then, we transmit, at the i^{th} channel use, the following coded bit z_i

$$z_i = \sum_{j=1}^{k} u_j g_{ji},$$

  where g_{ji} is the j^{th} element of $\mathbf{g}_i$. For each message $\mathbf{u}$, we send n coded symbols. Hence, the encoder is given by:

$$\begin{aligned} f_l : \mathcal{V}^k &\to \mathcal{V}^n \\ \mathbf{u} &\mapsto \mathbf{z} = f_l(\mathbf{u}) = \mathbf{u}^T \mathbf{G}_l. \end{aligned}$$

- The decoder decodes on the fly. Whenever it receives k linearly independent coded symbols, it decodes the message. Otherwise, it declares the packet to be erased.

7.4.3 Average Age of Random Codes

Fix the couple (n, k) and let $\mathcal{C}_1(n, k)$ be a given coding scheme generated as described in §7.4.2. We define $\Delta_{\epsilon,(n,k)}$ to be the expected average age of the coding scheme induced by a random linear (n, k)-scheme generated as above.

Definition 7.5. *For fixed (n, k),*

$$\Delta_{\epsilon,(n,k)} = \mathbb{E}_{\mathcal{C}(n,k)} \left(\Delta_{\epsilon,\mathcal{C}} \right), \tag{7.13}$$

where the expectation is taken over all random linear (n, k)-schemes $\mathcal{C}(n, k)$.

Due to the ergodicity of the system, almost surely, for any randomly generated (n, k)-scheme $\mathcal{C}$, we obtain $\Delta_{\epsilon,\mathcal{C}} = \Delta_{\epsilon,\mathcal{C}_1}$. Thus,

$$\Delta_{\epsilon,(n,k)} = \mathbb{E}_{\mathcal{C}(n,k)} \left(\Delta_{\epsilon,\mathcal{C}} \right) = \Delta_{\epsilon,\mathcal{C}_1}. \tag{7.14}$$

The contribution of the random coding argument in this context is the following: if we show that, for a given (n, k)-scheme $\mathcal{C}$, $\Delta_{\epsilon,\mathcal{C}} < \infty$, then $\Delta_{\epsilon,(n,k)} = \Delta_{\epsilon,\mathcal{C}} < \infty$ and there must exist a linear (n, k)-scheme $\mathcal{C}'$ such that $\Delta_{\epsilon,\mathcal{C}'} \leq \Delta_{\epsilon,(n,k)}$. Thus, the optimal average age $\Delta_\epsilon \leq \Delta_{\epsilon,\mathcal{C}'} \leq \Delta_{\epsilon,(n,k)}$ for all possible values of n[4]. Therefore,

$$\Delta_\epsilon \leq \min_{n \geq k} \Delta_{\epsilon,(n,k)}. \tag{7.15}$$

Equation (7.15) gives an upper bound on the optimal average age. In the rest of this chapter we will focus on characterizing this bound.

[3] In this chapter, we assume all vectors to be column vectors.

[4] The value of k is considered fixed since we assume we have no control over the source alphabet.

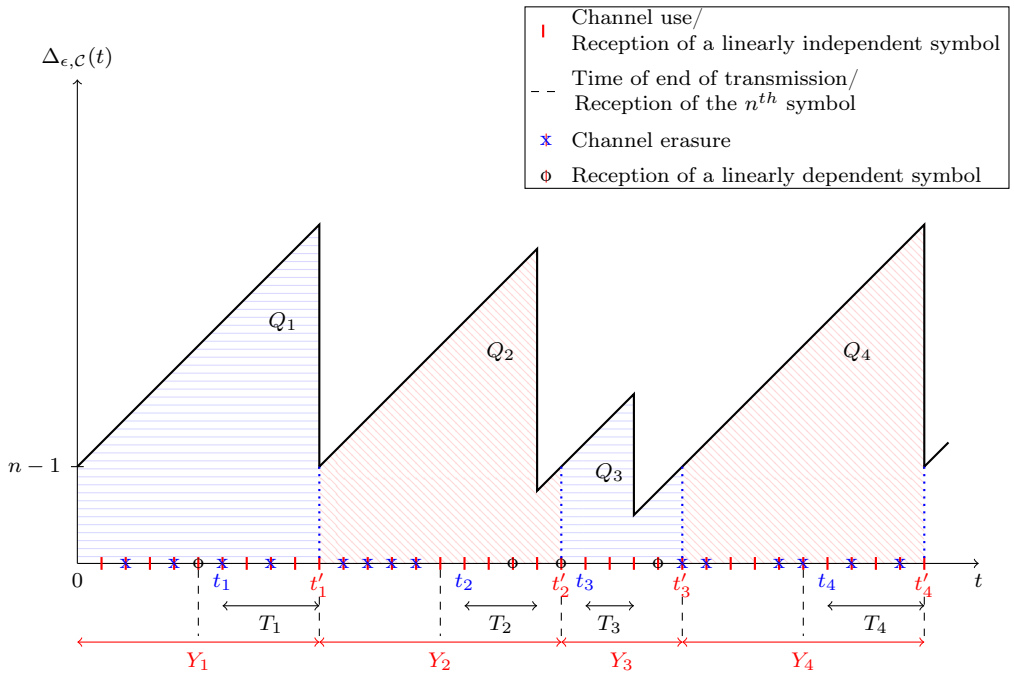


Figure 7.4 – Variation of the instantaneous age when using a random code $\mathcal{C}$ with $n = 5$, $k = 3$. We assume we begin observing after a successful reception. Since $\lambda = \mu = 1$ then the interval between channel uses is one second.

7.4.4 Exact Upper Bound on the Optimal Average Age

Preliminaries

Let $\mathcal{C}$ be a randomly generated (n, k)-scheme. Fig. 7.4 illustrates the variation of the instantaneous age $\Delta_{\epsilon,\mathcal{C}}(t)$ when $n = 5$ and $k = 3$. Without loss of generality, we assume we begin observing right after the reception of a successful packet. We denote by t_j the generation time of the j^{th} successful packet and by t'_j the end of transmission time of this packet. We notice that for the j^{th} successfully received message, the instantaneous age at the end of transmission is $n - 1$ since we assume the transmission to begin at the same time as a packet is generated. Thus, when the transmission of this successful packet ends at time t'_j after n channel uses, the age of this packet is $\Delta_{\epsilon,\mathcal{C}}(t'_j) = n - 1$.

In the scenario depicted in Fig. 7.4, the first packet $\mathbf{u}_1$ is generated and encoded into a codeword $\mathbf{z}_1 = (z_{11}, z_{12}, \cdots, z_{1n})$ of length $n = 5$ at time $t = 1$. At that same instant, z_{11}, the first symbol of $\mathbf{z}_1$, is sent and received at the monitor. Since it is the first symbol, z_{11} is linearly independent. At time $t = 2$, the coded symbol z_{12} is erased but the coded symbol z_{13}, which is linearly independent from z_{11}, is received at $t = 3$. The forth coded symbol is also erased and the last coded symbol z_{15} is received. However, since z_{15} is linearly dependent on the previously received symbols, namely z_{11} and z_{13}, the first packet $\mathbf{u}_1$ is declared erased by the decoder and $\Delta_{\epsilon,\mathcal{C}}(t)$ increases linearly by an additional $n = 5$ seconds. The packet generated at $t_1 = 6$ is a successful update since the monitor receives $k = 3$ linearly independent symbols. Therefore, $\Delta_{\epsilon,\mathcal{C}}(t)$ drops to $n - 1$ at $t'_1 = 10$. An interesting observation is that, for a

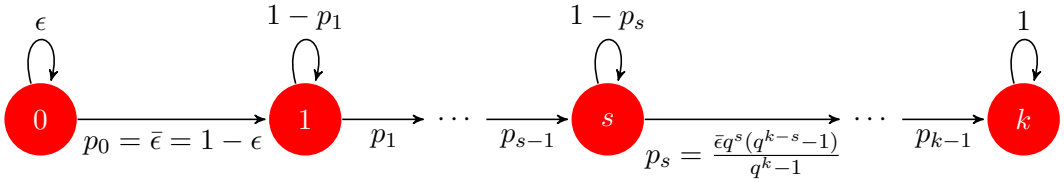


Figure 7.5 – Markov chain representing the dimension of a codeword at the receiver.

given successful packet, once k linearly independent coded symbols are received, any additional coded symbol is linearly dependent on them.

In this section we use a notation slightly different than the one introduced in Chapter 2:

- $Y_j = t'_j - t'_{j-1}$ is the interdeparture time between the j^{th} and $j-1^{th}$ successfully received updates.
- T_j is the number of channel uses between the decoding instant of the j^{th} successful packet and its generation time t_j.
- $R(\tau) = \max\{n : t'_n \leq \tau\}$ is the number of successfully received updates in the interval $[0, \tau]$.
- For a given packet i (not necessarily successful), B_i is the number of channel uses (or sent coded symbols) in order to receive exactly k linearly independent equations (coded symbols). Thus, a packet i is correctly decoded if $B_i \leq n$ for a fixed $n \geq k$.

Since the channel is memoryless and the different codes used in the scheme $\mathcal{C}$ are generated independently and in the same fashion, then the process $(B_i)_{i\geq 1}$ is i.i.d with a distribution identical to the random variable B.

The Distribution of B

Fig. 7.5 shows the Markov chain that represents the dimension, at the receiver, of the codeword relative to a certain update. The monitor receives the first coded symbol of a new codeword with probability $p_0 = \bar{\epsilon} = 1 - \epsilon$ and hence the dimension of this codeword at the receiver jumps to 1. If the first coded symbol is erased then the dimension of the codeword stays at 0. If the monitor has already received s linearly independent coded symbols, then it will receive the $s+1^{th}$ linearly independent coded symbol if: (i) The next transmitted coded symbol is not erased, and, (ii) the next transmitted coded symbol is linearly independent of all previously received symbols. Event (i) occurs with probability $\bar{\epsilon} = 1 - \epsilon$ and event (ii) happens with probability $\frac{q^s(q^{k-s}-1)}{q^k-1}$. Hence, for a given message, the dimension of its codeword at the receiver jumps from s to $s+1$ with probability

$$p_s = \frac{\bar{\epsilon}q^s(q^{k-s}-1)}{q^k-1},$$

where $0 \leq s \leq k-1$. If the next transmitted coded symbol is erased or linearly dependent on the previously received coded symbols, then the dimension of the codeword at the monitor stays s. As discussed before, once the monitor receives k linearly independent coded symbols, the dimension of the codeword stays at k and all subsequent coded symbols are linearly dependent on the previously non-erased coded symbols.

From the above description, we can deduce that B is the number of visits to the states $\{0, 1, \cdots, k-1\}$ before reaching state k for the first time.

Remark 7.1. *Since $p_s = \frac{\bar{\epsilon}(q^k - q^s)}{q^k - 1}$, then p_s is a decreasing function of s. This means that whenever the decoder receives a non-erased coded symbol that is linearly independent from all previously received coded symbols, and the system jumps to state s, then it becomes harder to receive a new linearly independent coded symbol. That's why, the system has a higher probability to spend more time in state s than in previous states.*

Definition 7.6. *Let L_s be the number of trials to pass from state s to state $s+1$ in Fig. 7.5, $0 \leq s \leq k-1$. L_s has a geometric distribution with success probability $p_s = \frac{\bar{\epsilon} q^s (q^{k-s} - 1)}{q^k - 1}$. Thus,*

$$\mathbb{P}(L_s = l) = (1 - p_s)^{l-1} p_s, \quad l = 1, 2, 3, \cdots$$

Corollary 7.1. *From Definition 7.6, we can write*

$$B = \sum_{s=0}^{k-1} L_s, \tag{7.16}$$

where the L_s's are independent.

Lemma 7.4. *The moment generating function of the random variable B is*

$$\phi_B(t) = \mathbb{E}\left(e^{tB}\right) = \left(\prod_{s=0}^{k-1} \left(q^k - q^s\right)\right) \left(\prod_{s=0}^{k-1} \frac{\bar{\epsilon} e^t}{q^k - 1 + e^t(1 - \epsilon q^k - \bar{\epsilon} q^s)}\right). \tag{7.17}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left(e^{tB}\right) = \mathbb{E}\left(e^{t\sum_{s=0}^{k-1} L_s}\right) &= \prod_{s=0}^{k-1} \mathbb{E}\left(e^{tL_s}\right) \\
&= \prod_{s=0}^{k-1} \sum_{l=1}^{\infty} e^{tl} (1 - p_s)^{l-1} p_s \\
&= \prod_{s=0}^{k-1} \frac{p_s e^t}{1 - (1 - p_s) e^t},
\end{aligned}$$

where the third equality is because the $(L_s)_{s=0}^{k-1}$ are mutually independent. Replacing p_s by its expression $p_s = \frac{\bar{\epsilon} q^s (q^{k-s} - 1)}{q^k - 1}$, we obtain (7.17). □

Corollary 7.2. *The expected value of B is*

$$\mathbb{E}(B) = \frac{q^k - 1}{1 - \epsilon} \sum_{s=0}^{k-1} \frac{1}{q^k - q^s}. \tag{7.18}$$

Proof.

$$\mathbb{E}(B) = \left. \frac{\mathrm{d}\phi_B(t)}{\mathrm{d}t} \right|_{t=0},$$

where $\phi_B(t)$ is given by (7.17). □

Remark 7.2. *The expected value of B can be also computed by using* (7.16)*:*

$$\mathbb{E}(B) = \sum_{s=0}^{k-1} \mathbb{E}(L_s) = \sum_{s=0}^{k-1} \frac{1}{p_s}.$$

Packet Erasure Probability

A packet or source symbol j is correctly received if for a given blocklength n, we have $B_j \leq n$. Otherwise, we declare the packet to be lost. Thus, we define the packet erasure probability ϵ_p to be

$$\epsilon_p = \mathbb{P}(B > n) = \sum_{l=n+1}^{\infty} \mathbb{P}(B = l), \tag{7.19}$$

where the distribution of B is given by Lemma 7.4. We call $1 - \epsilon_p = \mathbb{P}(B \leq n)$ to be the packet success probability.

The Age Analysis

Definition 7.7. *In every interdeparture interval Y_j, we call H_j the number of erased packets before the reception of a successful update. H_j is geometric with success probability ϵ_p, so*

$$\mathbb{P}(H_j = x) = \epsilon_p^x (1 - \epsilon_p), \quad x = 0, 1, 2, \cdots$$

We use Definition 7.7 to characterize the interdeparture interval. Indeed, any interdeparture interval is the sum of two components: The time spending unsuccessful packets followed by the service time of the successful update. Since each transmitted packet takes n channel uses and $\mu = 1$, then the j^{th} interdeparture time can be written as

$$Y_j = \frac{n}{\mu} H_j + \frac{n}{\mu} = n(H_j + 1), \quad j \geq 1. \tag{7.20}$$

Given that we assume a memoryless erasure channel and independently generated packets, then the process that consists of the number of erased packets $(H_j)_{j \geq 1}$ is i.i.d. Since the interdeparture interval Y_j is function of only H_j then the sequence $(Y_j)_{j \geq 1}$ is also i.i.d. Hence the following lemma:

Lemma 7.5. *The process $R(\tau) = \max\{n : t'_n \leq \tau\}$ is a renewal process with the interdeparture times $(Y_j)_{j\geq 1}$ being the renewal intervals.*

The importance of Lemma 7.5 stems from the fact that it shows that the instantaneous age process $\Delta_{\epsilon,\mathcal{C}}(t)$ is mean-ergodic (see Definition 2.1).

Lemma 7.6.

$$\Delta_{\epsilon,\mathcal{C}} = \lim_{\tau\to\infty} \frac{1}{\tau}\int_0^\tau \Delta_{\epsilon,\mathcal{C}}(t)\mathrm{d}t = \frac{\mathbb{E}(Q)}{\mathbb{E}(Y)}, \tag{7.21}$$

where Q is the steady-state counterpart of the shaded area Q_j shown in Fig. 7.4 and Y is the steady-state counterpart of the interdeparture interval Y_j.

Proof. We will use the DTA introduced in Section 2.3 to compute the average age. By Lemma 7.5, $R(\tau)$ forms a renewal process and hence by [58] we know that $\lim_{\tau\to\infty}\frac{R(\tau)}{\tau} = \frac{1}{\mathbb{E}(Y)}$, where Y is the steady-state interdeparture random variable. By defining $Q_j = \int_{t'_{j-1}}^{t'_j} \Delta_{\epsilon,\mathcal{C}}(t)\mathrm{d}t$ to be the reward function over the renewal period Y_j, we get using renewal reward theory [16, 58] that

$$\Delta_{\epsilon,\mathcal{C}} = \lim_{\tau\to\infty}\frac{1}{\tau}\int_0^\tau \Delta_{\epsilon,\mathcal{C}}(t)\mathrm{d}t = \lim_{\tau\to\infty}\frac{R(\tau)}{\tau}\frac{1}{R(\tau)}\sum_{j=1}^{R(\tau)} Q_j = \frac{\mathbb{E}(Q_j)}{\mathbb{E}(Y_j)} < \infty.$$

□

Before computing the average age, we still need one more lemma that gives the distribution of the random variable T_j, $j \geq 1$.

Lemma 7.7. *Let T be the steady-state counterpart of the number of channel uses T_j between the decoding instant of the j^{th} successful packet and its generation time t_j. Then,*

$$\mathbb{P}(T = x) = \frac{\mathbb{P}(B = x)\mathbb{1}_{\{x\leq n\}}}{\mathbb{P}(B \leq n)}, \tag{7.22}$$

where $\mathbb{1}_{\{.\}}$ is the indicator function.

Proof. A packet is successfully decoded if the decoder receives exactly k linearly independent coded symbols in less than n channel uses. Thus, for the j^{th} successful packet we have that

$$\mathbb{P}(T_j = x) = \mathbb{P}(B = x | B \leq n).$$

□

We are now ready to give the main theorem of this section.

Theorem 7.5. *Assume an EC(ϵ) and an (n, k)-coding scheme $\mathcal{C}$ as defined in Section 7.4.2. The average age $\Delta_{\epsilon,\mathcal{C}}$ corresponding to such setup is given by*

$$\Delta_{\epsilon,\mathcal{C}} = \mathbb{E}(T) - 1 + \frac{n(1+\epsilon_p)}{2(1-\epsilon_p)}, \tag{7.23}$$

where ϵ_p is the packet erasure probability given by (7.19).

Proof. From (7.21), we know that we need to compute $\mathbb{E}(Q)$ and $\mathbb{E}(Y)$. We start with $\mathbb{E}(Y)$. We have shown that for any $j \geq 1$, $Y_j = n(H_j + 1)$. Thus, in steady-state, we can write $Y = n(H + 1)$, where H is the steady-state counterpart of H_j. Hence,

$$\mathbb{E}(Y) = n(\mathbb{E}(H) + 1) = n\left(\frac{\epsilon_p}{1-\epsilon_p} + 1\right) = \frac{n}{1-\epsilon_p}, \tag{7.24}$$

where the second equality is due to the fact that H has a geometric distribution with success probability ϵ_p as seen in Definition 7.7.

Now we turn to $\mathbb{E}(Q)$. For any $j \geq 1$, the shaded area Q_j shown in Fig. 7.4 is the sum of the areas of two trapezoids: a large trapezoid with height $n(H_j + T_j)$ and a smaller one with height $n - T_j$. Thus,

$$\begin{aligned} Q_j &= \frac{(n-1+n-1+nH_j+T_j)(nH_j+T_j)}{2} + \frac{(T_j-1+n-1)(n-T_j)}{2} \\ &= \frac{1}{2}\left(2n(n-1)H_j + 2nT_j(1+H_j) + n^2H_j^2 + n(n-2)\right). \end{aligned}$$

Note that H_j and T_j are independent. Therefore,

$$\begin{aligned} \mathbb{E}(Q_j) &= \frac{1}{2}\left(\mathbb{E}\left(2n(n-1)H_j + 2nT_j(1+H_j) + n^2H_j^2 + n(n-2)\right)\right) \\ &= n(n-1)\mathbb{E}(H_j) + n\mathbb{E}(T_j(1+H_j)) + \frac{n^2\mathbb{E}\left(H_j^2\right)}{2} + \frac{n(n-2)}{2} \\ &= n(n-1)\mathbb{E}(H_j) + n\mathbb{E}(T_j)\mathbb{E}(1+H_j) + \frac{n^2\mathbb{E}\left(H_j^2\right)}{2} + \frac{n(n-2)}{2}. \end{aligned}$$

In steady-state, we obtain

$$\mathbb{E}(Q) = n(n-1)\mathbb{E}(H) + n\mathbb{E}(T)\mathbb{E}(1+H) + \frac{n^2\mathbb{E}\left(H^2\right)}{2} + \frac{n(n-2)}{2}. \tag{7.25}$$

Replacing $\mathbb{E}(Y)$ and $\mathbb{E}(Q)$ in (7.21) by their expressions in (7.24) and (7.25), we obtain (7.23). □

In the expression of $\Delta_{\epsilon,\mathcal{C}}$, $\mathbb{E}(T)$ and ϵ_p cannot be easily expressed in function of ϵ, k and n. That is why we study $\Delta_{\epsilon,\mathcal{C}}$ in the next two subsections by presenting upper and lower bounds on the expression in (7.23).

7.4.5 Bounding $\Delta_{\epsilon,\mathcal{C}}$

Definition 7.8. *We define $\tilde{B}$ to be the sum of k i.i.d random variables distributed like L_0. We also define $\hat{B}$ to be the sum of k i.i.d random variables distributed like L_{k-1}. Formally,*

$$\tilde{B} = \sum_{i=0}^{k-1} L_0^{(i)} \quad \textit{and} \quad \hat{B} = \sum_{i=0}^{k-1} L_{k-1}^{(i)}, \tag{7.26}$$

where L_0 is geometrically distributed with success probability $\bar{\epsilon} = 1 - \epsilon$ and L_{k-1} is also geometrically distributed with success probability $p_{k-1} = \frac{\bar{\epsilon}q^{k-1}(q-1)}{q^k-1}$.

Lemma 7.8. *The random variables $\tilde{B}$ and $\hat{B}$ defined in Definition 7.8 are both negative binomials with*

$$\mathbb{P}(\tilde{B} = x) = \binom{x-1}{k-1}(1-\epsilon)^k \epsilon^{x-k},$$

and

$$\mathbb{P}(\hat{B} = x) = \binom{x-1}{k-1}(p_{k-1})^k (1-p_{k-1})^{x-k},$$

where $x = k, k+1, k+2,$ etc.

Proof. $\tilde{B}$ is the sum of k i.i.d geometric random variables with success probability $1-\epsilon$. Similarly, $\hat{B}$ is the sum of k i.i.d geometric random variables with success probability p_{k-1}. □

Lemma 7.9. *Given $B = \sum_{s=0}^{k-1} L_s$ and $\tilde{B}$ and $\hat{B}$ as defined in Definition 7.8, the following relations hold for $n \geq k$:*

1. $\mathbb{P}(\tilde{B} \leq n) \geq \mathbb{P}(B \leq n)$,
2. $\mathbb{E}(\tilde{B}) \leq \mathbb{E}(B)$,
3. $\mathbb{P}(\hat{B} \leq n) \leq \mathbb{P}(B \leq n)$.
4. $\mathbb{E}(\hat{B}) \geq \mathbb{E}(B)$,

Proof. We use a coupling argument to prove these relations. We first start by proving the first two identities. First notice that the probabilities $p_s = \frac{\bar{\epsilon} q^s (q^{k-s}-1)}{q^k - 1}$ are decreasing in s, where $0 \leq s \leq k-1$. This means that

$$p_0 \geq p_1 \geq \cdots \geq p_{k-1}.$$

Let $\tilde{B} = \sum_{s=0}^{k-1} L_0^{(s)}$, with the $(L_0^{(s)})_{s=0}^{k-1}$ being i.i.d and similarly distributed to L_0. Define the random variables $A_s = L_0^{(s)} + J_s$ with $L_0^{(s)}$ and J_s independent and J_s distributed as follows, for $0 \leq s \leq k-1$:

$$\mathbb{P}(J_s = x) = \begin{cases} \frac{p_s}{p_0} & \text{if } x = 0 \\ \frac{p_0 - p_s}{p_0}(1-p_s)^{x-1} p_s & \text{if } x = 1, 2, 3 \cdots \\ 0 & \text{else.} \end{cases}$$

Therefore, the distribution of A_s, for $0 \leq s \leq k-1$ and for $x \geq 1$, is

$$\begin{aligned}
\mathbb{P}(A_s = x) &= \mathbb{P}(L_0^{(s)} + J_s = x) \\
&= \mathbb{P}(J_s = x - L_0^{(s)}) \\
&= \sum_{l=1}^{x} \mathbb{P}(L_0^{(s)} = l)\mathbb{P}(J_s = x - l | L_0^{(s)} = l) \\
&= \sum_{l=1}^{x} (1-p_0)^{l-1} p_0 \left(\frac{p_s}{p_0} \mathbb{1}_{\{x-l=0\}} + \frac{p_0 - p_s}{p_0} (1-p_s)^{x-l-1} p_s \mathbb{1}_{\{x-l\neq 0\}} \right) \\
&= p_s(1-p_0)^{x-1} + p_s(p_0 - p_s)(1-p_s)^{x-2} \sum_{l=1}^{x-1} \left(\frac{1-p_0}{1-p_s} \right)^{l-1} \\
&= p_s(1-p_0)^{x-1} + p_s(p_0 - p_s)(1-p_s)^{x-1} \left(\frac{1 - \left(\frac{1-p_0}{1-p_s} \right)^{x-1}}{p_0 - p_s} \right) \\
&= p_s(1-p_s)^{x-1}.
\end{aligned}$$

This means that, for any $s \in \{0, 1, \cdots, k-1\}$, A_s is distributed similarly to L_s: They are both geometric random variables with success probability p_s. Since the sequence $(L_0^{(s)})_{s=0}^{k-1}$ is i.i.d and, for any s, the variable J_s is independent of $L_0^{(s)}$ and independent of J_i, $i \neq s$, then, for any $s \neq i$, A_s and A_i are independent. Define

$$O = \sum_{s=0}^{k-1} A_s = \tilde{B} + \sum_{s=0}^{k-1} J_s.$$

Then, O and $B = \sum_{s=0}^{k-1} L_s$ are identically distributed, i.e. $\mathbb{P}(O = x) = \mathbb{P}(B = x)$ $\forall\ x \geq 1$. Using this fact and noticing that $O \geq \tilde{B}$ with probability 1, we deduce that the event $\{O \leq n\}$ is a subset of the event $\{\tilde{B} \leq n\}$. Hence,

$$\mathbb{P}(B \leq n) = \mathbb{P}(O \leq n) \leq \mathbb{P}(\tilde{B} \leq n).$$

This inequality also implies $\mathbb{P}(B \geq n) \geq \mathbb{P}(\tilde{B} \geq n)$. To prove item 2., we first notice that

$$\mathbb{E}(O) = \mathbb{E}(\tilde{B}) + \sum_{s=0}^{k-1} \mathbb{E}(J_s).$$

Since $J_s \geq 0$ for any s, then $\mathbb{E}(O) \geq \mathbb{E}(\tilde{B})$. Given that O and B are identically distributed, then

$$\mathbb{E}(O) = \mathbb{E}(B) \geq \mathbb{E}(\tilde{B}).$$

To prove items 3. and 4. of Lemma 7.9 we use a similar argument as the one used for items 1. and 2.. However, here, we define $A_s = L_s + J_s$ with L_s and J_s independent and J_s distributed as follows, for $0 \leq s \leq k-1$:

$$\mathbb{P}(J_s = x) = \begin{cases} \frac{p_{k-1}}{p_s} & \text{if } x = 0 \\ \frac{p_s - p_{k-1}}{p_s}(1-p_{k-1})^{x-1} p_{k-1} & \text{if } x = 1, 2, 3 \cdots \\ 0 & \text{else.} \end{cases}$$

In this case, for any s, A_s has the same distribution as L_{k-1}. Since $(L_s)_{s=1}^{k}$ are independent and, for any s, J_s is independent of L_s then, for any $s \neq i$, A_s and A_i are independent and identically distributed. Define

$$O = \sum_{s=0}^{k-1} A_s = B + \sum_{s=0}^{k-1} J_s.$$

Then, O and $\hat{B} = \sum_{s=0}^{k-1} L_{k-1}^{(s)}$ are identically distributed, i.e. $\mathbb{P}(O = x) = \mathbb{P}(\hat{B} = x)$ $\forall\, x \geq 1$. Using this fact and noticing that $O \geq B$ with probability 1, we deduce that the event $\{O \leq n\}$ is a subset of the event $\{B \leq n\}$. Hence,

$$\mathbb{P}(\hat{B} \leq n) = \mathbb{P}(O \leq n) \leq \mathbb{P}(B \leq n).$$

This inequality also implies $\mathbb{P}(\hat{B} \geq n) \geq \mathbb{P}(B \geq n)$. To prove item 4., we first notice that

$$\mathbb{E}(O) = \mathbb{E}(B) + \sum_{s=0}^{k-1} \mathbb{E}(J_s).$$

Since $J_s \geq 0$ for any s, then $\mathbb{E}(O) \geq \mathbb{E}(B)$. Given that O and $\hat{B}$ are identically distributed, then

$$\mathbb{E}(O) = \mathbb{E}(\hat{B}) \geq \mathbb{E}(B).$$

□

Lemma 7.9 can be interpreted as follows: $\tilde{B}$ can be seen as the number of channel uses in order to receive exactly k linearly independent coded symbols when any k coded symbols are linearly independent. This means that $\tilde{B}$ corresponds to the number of channel uses needed to decode a packet when the encoders of the (n, k)-scheme only use MDS codes. Hence, $\tilde{B}$ is equivalent to the number of channel uses needed to receive exactly k non-erased coded symbols. Intuitively, we would expect to need a number $\tilde{B}$ of channel uses to receive k non-erased coded symbols which is smaller than the number B needed to receive k linearly independent coded symbols. This explains the intuition behind items 1. and 2. in Lemma 7.9. On the opposite side of the spectrum, $\hat{B}$ can be seen as a worst case scenario since the jump from state s to state $s+1$ in Fig. 7.5 occurs with the smallest possible probability, namely p_{k-1}. This discussion leads us to the idea that $\Delta_{\epsilon,\mathcal{C}}$ could be upper bounded by the average age corresponding to a coding system with $\hat{B}$ as the number of channel uses need to receive exactly k linearly independent coded symbols. Similarly, $\Delta_{\epsilon,\mathcal{C}}$ could be lower bounded by the average age achieved using only MDS codes with $\tilde{B}$ as the number of channel uses needed to receive k linearly independent coded symbols.

Upper Bound on $\Delta_{\epsilon,\mathcal{C}}$

Recall from Theorem 7.5 that

$$\Delta_{\epsilon,\mathcal{C}} = \mathbb{E}(T) - 1 + \frac{n(1+\epsilon_p)}{2(1-\epsilon_p)}.$$

Based on Lemma 7.9 and Lemma 7.7, we can write

$$\begin{aligned}\mathbb{E}(T) = \sum_{x=k}^{n} x\frac{\mathbb{P}(B=x)}{\mathbb{P}(B\leq n)} &= \frac{\mathbb{E}(B\mathbb{1}_{\{B\leq n\}})}{\mathbb{P}(B\leq n)} \\ &\leq \frac{\mathbb{E}(\hat{B}\mathbb{1}_{\{\hat{B}\leq n\}})}{\mathbb{P}(\hat{B}\leq n)} \\ &= \sum_{x=k}^{n} x\frac{\mathbb{P}(\hat{B}=x)}{\mathbb{P}(\hat{B}\leq n)}. \end{aligned} \tag{7.27}$$

From Lemma 7.8, we know that $\hat{B}$ is a negative binomial random variable. Hence the bound in (7.27) becomes

$$\begin{aligned}\mathbb{E}(T) &\leq \sum_{x=k}^{n} x\frac{\mathbb{P}(\hat{B}=x)}{\mathbb{P}(\hat{B}\leq n)} \\ &= \sum_{x=k}^{n} x\frac{\binom{x-1}{k-1}(1-p_{k-1})^{x-k}{p_{k-1}}^{k}}{\mathbb{P}(\hat{B}\leq n)} \\ &= \frac{k}{\mathbb{P}(\hat{B}\leq n)}\sum_{x=k}^{n}\binom{x}{k}(1-p_{k-1})^{x-k}(p_{k-1})^{k}. \end{aligned} \tag{7.28}$$

Let $\hat{\hat{B}} = \sum_{i=0}^{k} L_{k-1}^{(i)}$, where $(L_{k-1}^{(i)})_{i=0}^{k}$ are i.i.d with a marginal distribution identical to L_{k-1}. Hence $\hat{\hat{B}}$ is also a negative binomial and

$$\mathbb{P}(\hat{\hat{B}}=x) = \binom{x-1}{k}(1-p_{k-1})^{x-k-1}(p_{k-1})^{k+1}, \quad x\geq k+1.$$

We use the same trick as in [76] and set $x' = x+1$ in (7.28). This leads to

$$\mathbb{E}(T) \leq \frac{k}{\mathbb{P}(\hat{B}\leq n)}\sum_{x'=k+1}^{n+1}\binom{x'-1}{k}(1-p_{k-1})^{x'-k-1}(p_{k-1})^{k} = \frac{k\mathbb{P}(\hat{\hat{B}}\leq n+1)}{p_{k-1}\mathbb{P}(\hat{B}\leq n)}, \tag{7.29}$$

where $\mathbb{P}(\hat{\hat{B}}\leq n+1) = \sum_{x=k+1}^{n+1}\binom{x-1}{k}(1-p_{k-1})^{x-k-1}(p_{k-1})^{k+1}$. In addition to that, we know from Lemma 7.9 that

$$\epsilon_p = \mathbb{P}(B\geq n+1) \leq \mathbb{P}(\hat{B}\geq n+1).$$

This means that $\frac{1+\epsilon_p}{1-\epsilon_p} \leq \frac{1+\mathbb{P}(\hat{B}\geq n+1)}{1-\mathbb{P}(\hat{B}\geq n+1)}$. Using this result and (7.29) on $\Delta_{\epsilon,\mathcal{C}}$, we get

$$\begin{aligned}\Delta_{\epsilon,\mathcal{C}} &\leq \frac{k\mathbb{P}(\hat{\hat{B}}\leq n+1)}{p_{k-1}\mathbb{P}(\hat{B}\leq n)} - 1 + \frac{n\left(1+\mathbb{P}(\hat{B}\geq n+1)\right)}{2\left(1-\mathbb{P}(\hat{B}\geq n+1)\right)} \\ &= \frac{2np_{k-1} - p_{k-1}\mathbb{P}(\hat{B}\leq n)(n+2) + 2k\mathbb{P}(\hat{\hat{B}}\leq n+1)}{2p_{k-1}\mathbb{P}(\hat{B}\leq n)}, \end{aligned}$$

where the second equality is obtained by using $\mathbb{P}(\hat{B} \geq n+1) = 1 - \mathbb{P}(\hat{B} \leq n)$. We denote by $\Delta^{ub}_{\epsilon,\mathcal{C}}$ the upper we just found. Thus,

$$\Delta^{ub}_{\epsilon,\mathcal{C}} = \frac{2np_{k-1} - p_{k-1}\mathbb{P}(\hat{B} \leq n)(n+2) + 2k\mathbb{P}(\hat{\hat{B}} \leq n+1)}{2p_{k-1}\mathbb{P}(\hat{B} \leq n)}. \tag{7.30}$$

Lower Bound on $\Delta_{\epsilon,\mathcal{C}}$

Let $\tilde{\tilde{B}} = \sum_{i=0}^{k} L_0^{(i)}$, where $(L_0^{(i)})_{i=0}^{k}$ are i.i.d with a marginal distribution identical to L_0. Hence $\tilde{\tilde{B}}$ is also a negative binomial and

$$\mathbb{P}(\tilde{\tilde{B}} = x) = \binom{x-1}{k} \epsilon^{x-k-1}(1-\epsilon)^{k+1}, \quad x \geq k+1.$$

Using Lemma 7.9 and an argument identical to that used for the computation of the upper bound $\Delta^{ub}_{\epsilon,\mathcal{C}}$ we show that $\Delta_{\epsilon,\mathcal{C}} \geq \Delta^{lb}_{\epsilon,(n,k)}$, where

$$\Delta^{lb}_{\epsilon,(n,k)} = \frac{2n(1-\epsilon) - (1-\epsilon)\mathbb{P}(\tilde{B} \leq n)(n+2) + 2k\mathbb{P}(\tilde{\tilde{B}} \leq n+1)}{2(1-\epsilon)\mathbb{P}(\tilde{B} \leq n)}. \tag{7.31}$$

Remark 7.3. *The lower bound found here is similar to the average age derived in [76] for the finite redundancy (FR) case. However, the time scale is different since Yates et al. in [76] assume the source generates a new update at the same instant it finishes transmitting the previous one. Whereas in our case we assume we generate and begin transmitting a new packet $\frac{1}{\mu}$ seconds after the last update finishes transmission.*

7.4.6 Age-Optimal Codes

We have already discussed that the lower bound on $\Delta_{\epsilon,\mathcal{C}}$, $\Delta^{lb}_{\epsilon,(n,k)}$, corresponds to the average age when the (n,k)-scheme uses only MDS codes with $\tilde{B}$ as the number of channel uses needed to receive k linearly independent coded symbols. Recall from Lemma 7.3 that, for a given couple (n,k), using an MDS code is optimal. This observation gives a different explanation on why the expression found in (7.31) is indeed a lower bound on the average age corresponding to a scheme using any other type of codes than MDS, in particular a code generated randomly. This means that the lower bound is universal over all codes and the optimal achievable age

$$\Delta^{ub}_{\epsilon,\mathcal{C}} \geq \Delta_{\epsilon,\mathcal{C}} \geq \Delta_{\epsilon} \geq \min_{n \geq k} \Delta^{lb}_{\epsilon,(n,k)}, \tag{7.32}$$

where $\mathcal{C}$ is an (n,k)-random code. However, for a given (n,k), an explicit construction of an MDS code is not always available. In this section, we show that if the channel-input alphabet is large enough, then random codes are age-optimal.

Theorem 7.6. *Fix a couple (n,k). We have that $\forall\ \delta > 0$, $\exists q_0 > 0$ such that $\forall\ q \geq q_0$, there exists an (n,k)-random code $\mathcal{C}$ such that*

$$|\Delta_{\epsilon,\mathcal{C}} - \Delta^{lb}_{\epsilon,(n,k)}| < \delta. \tag{7.33}$$

This means that for a channel-input alphabet large enough (q large), random codes are age-optimal and

$$\Delta_\epsilon \doteq \min_{n\geq k} \Delta_{\epsilon,\mathcal{C}}, \tag{7.34}$$

where $\mathcal{C}$ is an (n,k)-random code and the dot above the equal sign refers to the fact that the difference between the two sides approaches zero as q gets large.

Proof. For a given random code $\mathcal{C}$, recall that

$$\Delta_{\epsilon,\mathcal{C}} = \mathbb{E}(T) - 1 + \frac{n(1+\epsilon_p)}{2(1-\epsilon_p)}.$$

We notice that $\mathbb{E}(T)$ and ϵ_p both depend only on the distribution of $B = \sum_{s=0}^{k-1} L_s$. However, for any $s \in \{0, 1, \cdots, k-1\}$,

$$\lim_{q\to\infty} p_s = \lim_{q\to\infty} (1-\epsilon)\frac{q^k - q^s}{q^k - 1} = 1 - \epsilon = p_0.$$

This means that, for any s, L_s converges in distribution to L_0 as $q \to \infty$. Therefore, B converges in distribution to $\tilde{B} = \sum_{s=0}^{k-1} L_0^{(s)}$, as $q \to \infty$. Hence, as $q \to \infty$, $\Delta_{\epsilon,\mathcal{C}}$ converges to $\Delta^{lb}_{\epsilon,(n,k)}$. So, for q large enough, we can write

$$\Delta_{\epsilon,\mathcal{C}} = \Delta_{\epsilon,(n,k)} \doteq \Delta^{lb}_{\epsilon,(n,k)}.$$

From (7.15), we know that the optimal age, for a given q, is $\Delta_\epsilon \leq \min_{n\geq k} \Delta_{\epsilon,(n,k)}$. For large enough q, we have $\Delta_{\epsilon,(n,k)} \doteq \Delta^{lb}_{\epsilon,(n,k)}$. This means that asymptotically, $\Delta_\epsilon \dot{\leq} \min_{n\geq k} \Delta^{lb}_{\epsilon,(n,k)}$. However, from (7.32), we have that $\Delta_\epsilon \geq \min_{n\geq k} \Delta^{lb}_{\epsilon,(n,k)}$ for any q. Therefore, asymptotically

$$\Delta_\epsilon \doteq \min_{n\geq k} \Delta^{lb}_{\epsilon,(n,k)}.$$

□

7.4.7 Other Bounds and Approximations

Upper Bounding the Lower Bound

In Remark 7.3, we discussed how the lower bound found in (7.31) is similar, up to a time scale difference, to the average age computed by Yates et al. in [76, Section 3]. In this paper, the authors present a tight upper bound on the computed average age. We borrow the same techniques as in [76, Section 3.A] to upper bound $\Delta^{lb}_{\epsilon,(n,k)}$. Interestingly, simulations will show that the upper bound to $\Delta^{lb}_{\epsilon,(n,k)}$ is a tight approximation to $\Delta_{\epsilon,\mathcal{C}}$, the average age achieved when using a (n,k)-random code $\mathcal{C}$.

Recall that

$$\begin{aligned}\Delta^{lb}_{\epsilon,(n,k)} &= \frac{2n(1-\epsilon) - (1-\epsilon)\mathbb{P}(\tilde{B} \leq n)(n+2) + 2k\mathbb{P}(\tilde{\tilde{B}} \leq n+1)}{2(1-\epsilon)\mathbb{P}(\tilde{B} \leq n)} \\ &= \frac{k\mathbb{P}(\tilde{\tilde{B}} \leq n+1)}{(1-\epsilon)\mathbb{P}(\tilde{B} \leq n)} - 1 + \frac{n(2 - \mathbb{P}(\tilde{B} \leq n))}{2\mathbb{P}(\tilde{B} \leq n)}.\end{aligned} \tag{7.35}$$

Denote by $\tilde{\mu}_n = \frac{k\mathbb{P}(\tilde{\tilde{B}}\leq n+1)}{(1-\epsilon)\mathbb{P}(\tilde{B}\leq n)}$. From [76, Lemma 1], we know that $\tilde{\mu}_n \leq \min\left(n, \frac{k}{1-\epsilon}\right)$. Hence,

$$\Delta^{lb}_{\epsilon,(n,k)} \leq \frac{k}{1-\epsilon} - 1 + \frac{n(2-\mathbb{P}(\tilde{B}\leq n))}{2\mathbb{P}(\tilde{B}\leq n)}.$$

We denote by $\Delta^{*}_{\epsilon,(n,k)}$ this approximation. Thus,

$$\Delta^{*}_{\epsilon,(n,k)} = \frac{k}{1-\epsilon} - 1 + \frac{n(2-\mathbb{P}(\tilde{B}\leq n))}{2\mathbb{P}(\tilde{B}\leq n)}. \tag{7.36}$$

Remark 7.4. *We can apply the techniques discussed in [76, Section 3.A] in order to approximate the optimal codeword length n for $\Delta^{lb}_{\epsilon,(n,k)}$ and write $\Delta^{*}_{\epsilon,(n,k)}$ solely in function of ϵ, k, n and the size q of the channel-input alphabet.*

Another Upper Bound on $\Delta_{\epsilon,\mathcal{C}}$

We derive here a second upper bound on $\Delta_{\epsilon,\mathcal{C}}$ which is easier to compute than $\Delta^{ub}_{\epsilon,\mathcal{C}}$. First recall from Theorem 7.5 that

$$\Delta_{\epsilon,\mathcal{C}} = \mathbb{E}(T) - 1 + \frac{n(1+\epsilon_p)}{2(1-\epsilon_p)}.$$

However,

$$\begin{aligned}
\mathbb{E}(T) &= \sum_{x=k}^{n} x\frac{\mathbb{P}(B=x)}{\mathbb{P}(B\leq n)} \\
&= \frac{1}{\mathbb{P}(B\leq n)}\left(\sum_{x=k}^{\infty} x\mathbb{P}(B=x) - \sum_{x=n+1}^{\infty} x\mathbb{P}(B=x)\right) \\
&= \frac{1}{\mathbb{P}(B\leq n)}\left(\mathbb{E}(B) - \sum_{x=n+1}^{\infty} x\mathbb{P}(B=x)\right) \\
&\leq \frac{1}{\mathbb{P}(B\leq n)}\left(\mathbb{E}(B) - (n+1)(1-\mathbb{P}(B\leq n))\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
\Delta_{\epsilon,\mathcal{C}} &\leq \frac{1}{\mathbb{P}(B\leq n)}\left(\mathbb{E}(B) - (n+1)(1-\mathbb{P}(B\leq n))\right) - 1 + \frac{n(1+\epsilon_p)}{2(1-\epsilon_p)} \\
&= \frac{1}{1-\epsilon_p}\left(\mathbb{E}(B) - (n+1)\epsilon_p)\right) - 1 + \frac{n(1+\epsilon_p)}{2(1-\epsilon_p)} \\
&= \frac{\mathbb{E}(B)-1}{1-\epsilon_p} + \frac{n}{2}.
\end{aligned} \tag{7.37}$$

Whereas $\mathbb{E}(B)$ in (7.18) is easy to compute, $\epsilon_p = \mathbb{P}(B \geq n+1)$ is hard to compute due to the complex nature of the distribution of B (given in Lemma 7.4). To solve this problem, we use $\hat{B}$ as defined in Definition 7.8 to upper bound ϵ_p. Indeed, from Lemma 7.9 we know that

$$\epsilon_p = \mathbb{P}(B \geq n+1) \leq \mathbb{P}(\hat{B} \geq n+1).$$

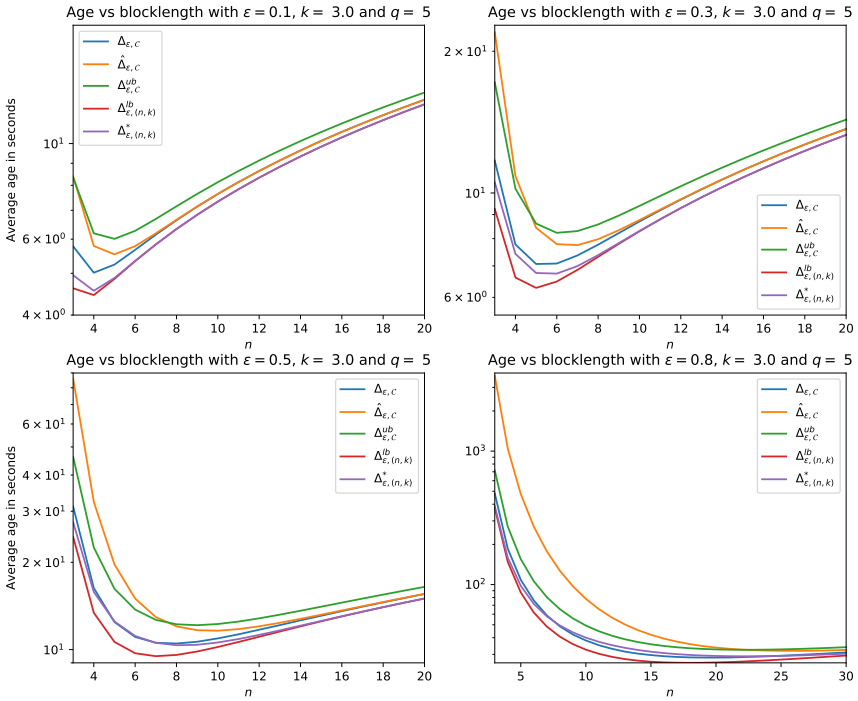


Figure 7.6 – Bounds on $\Delta_{\epsilon,\mathcal{C}}$ with respect to the blocklength n, with $k = 3$ and a channel-input alphabet of size $q = 5$. The age is in log scale.

Hence,

$$\Delta_{\epsilon,\mathcal{C}} \leq \frac{\mathbb{E}(B) - 1}{\mathbb{P}(\hat{B} \leq n)} + \frac{n}{2}.$$

Therefore, using (7.18), the new upper bound $\hat{\Delta}_{\epsilon,\mathcal{C}}$ is

$$\begin{aligned}
\hat{\Delta}_{\epsilon,\mathcal{C}} &= \frac{\mathbb{E}(B) - 1}{\mathbb{P}(\hat{B} \leq n)} + \frac{n}{2} \\
&= \frac{-1 + \frac{q^k - 1}{1-\epsilon} \sum_{s=0}^{k-1} \frac{1}{q^k - q^s}}{\mathbb{P}(\hat{B} \leq n)} + \frac{n}{2} \\
&= \frac{\epsilon + \left(q^k - 1\right) \sum_{s=1}^{k-1} \left(q^k - q^s\right)^{-1}}{(1-\epsilon)\mathbb{P}(\hat{B} \leq n)} + \frac{n}{2}.
\end{aligned} \tag{7.38}$$

7.4.8 Numerical Results

Fig. 7.6 and Fig. 7.8a corresponds to a system with a coding scheme where $k = 3$, $q = |\mathcal{V}| = 5$ and using a (n, k)-random code $\mathcal{C}$. Fig. 7.6 plots $\Delta_{\epsilon,\mathcal{C}}$ as well as the bounds and the approximation derived in Sections 7.4.5 and 7.4.7 with respect to the blocklength n, for four erasure channels with erasure probabilities $0.1, 0.3, 0.5, 0.8$. The tightness of the bounds with respect to $\Delta_{\epsilon,\mathcal{C}}$ differs according to the erasure probability:

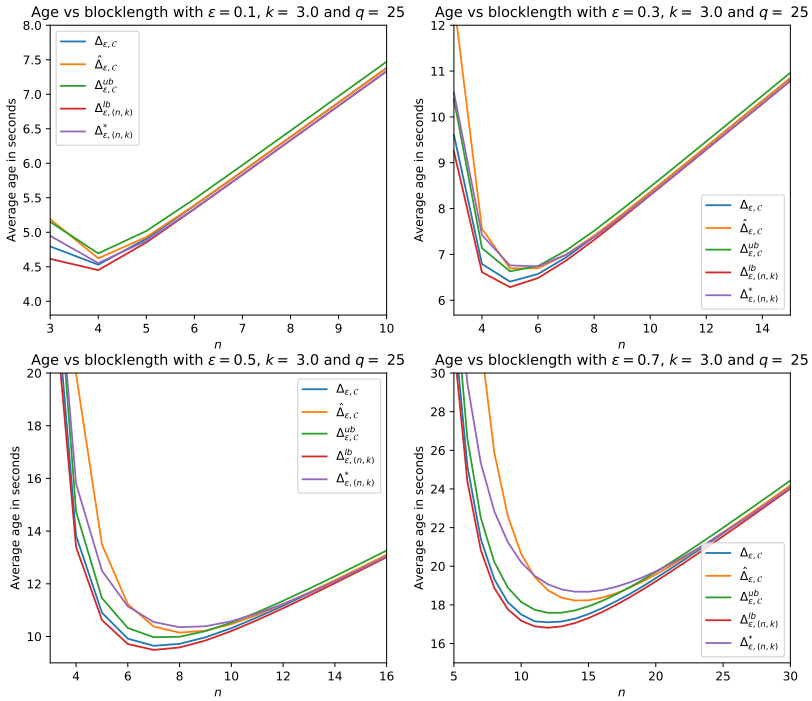


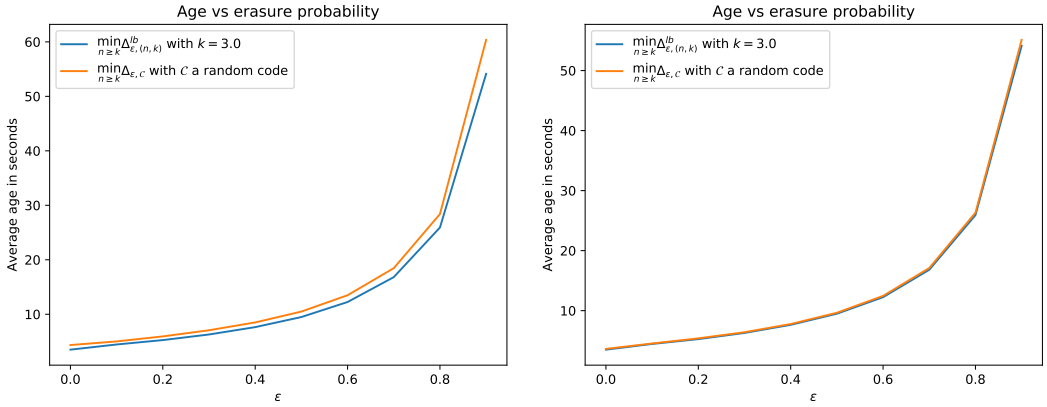
Figure 7.7 – Bounds on $\Delta_{\epsilon,\mathcal{C}}$ with respect to the blocklength n, with $k = 3$ and a channel-input alphabet of size $q = 25$.

- For all error probabilities, we notice that the upper bound $\hat{\Delta}_{\epsilon,\mathcal{C}}$ (the orange curve) is very tight (almost equal to $\Delta_{\epsilon,\mathcal{C}}$) at large enough n. However, the value n^* of the blocklength n starting which $\hat{\Delta}_{\epsilon,\mathcal{C}}$ becomes tight depends on ϵ: The larger the erasure probability, the larger the blocklength n. For instance, for $\epsilon = 0.1$ we have $n^* = 7$. But for $\epsilon = 0.5$, $n^* = 12$ and for $\epsilon = 0.8$ we have $n^* = 30$. For $n > n^*$, the upper bound $\hat{\Delta}_{\epsilon,\mathcal{C}}$ is tighter than all other bounds. Notice that for any n and any ϵ, $\Delta_{\epsilon,\mathcal{C}} \leq \hat{\Delta}_{\epsilon,\mathcal{C}}$.

- For the approximation $\Delta^*_{\epsilon,(n,k)}$, we notice that it becomes tighter as the erasure probability becomes larger. This is true especially at low values of n, more particularly for $n < n^*$. For this range of blocklength values the approximation $\Delta^*_{\epsilon,(n,k)}$ is the extremely close to $\Delta_{\epsilon,\mathcal{C}}$.

- For any value of n and any value of ϵ we observe that $\Delta^{lb}_{\epsilon,(n,k)} \leq \Delta_{\epsilon,\mathcal{C}}$ and $\Delta^{lb}_{\epsilon,(n,k)} \leq \Delta^*_{\epsilon,(n,k)}$. We notice that, for all values of ϵ, $\Delta^{lb}_{\epsilon,(n,k)}$ is close to $\Delta_{\epsilon,\mathcal{C}}$ at large n. Whereas, for small values of n, this lower bound does not show any noticeable behavioral modification as ϵ increases.

- The upper bound $\Delta^{ub}_{\epsilon,\mathcal{C}}$ is always larger than $\Delta_{\epsilon,\mathcal{C}}$. Even though at $n > n^*$ we observe that $\hat{\Delta}_{\epsilon,\mathcal{C}} \leq \Delta^{ub}_{\epsilon,\mathcal{C}}$, for $n \leq n^*$ the upper bound $\Delta^{ub}_{\epsilon,\mathcal{C}}$ is closer to $\Delta_{\epsilon,\mathcal{C}}$ than $\hat{\Delta}_{\epsilon,\mathcal{C}}$. In fact, as ϵ increases the gap between the two upper bounds increases also.



(a) $q = 5$. (b) $q = 25$.

Figure 7.8 – Bounds on the optimal achievable age Δ_ϵ with $k = 3$.

Fig. 7.6 also suggests that there exists, for each erasure probability, an optimal blocklength that minimizes $\Delta_{\epsilon,\mathcal{C}}$. This echoes the observations presented in Chapter 6 and in [76]. Moreover, each bound also has its optimal blocklength. Although the channel-input alphabet chosen is small ($k = 3$ and $q = 5$), we remark that the gap between $\Delta_{\epsilon,\mathcal{C}}$ and the lower bound $\Delta^{lb}_{\epsilon,(n,k)}$ is not too great irrespective of the value of ϵ. This means that even for small channel-input alphabets, the performance of the optimal linear code is not too far from the performance achieved by random coding. This idea is illustrated in Fig. 7.8a. In this last figure, we find and plot, at each value of ϵ, the minimum (with respect to n) of $\Delta_{\epsilon,\mathcal{C}}$ and $\Delta^{lb}_{\epsilon,(n,k)}$. We observe that these two minimums are close to each other. Since

$$\min_{n\geq k} \Delta^{lb}_{\epsilon,(n,k)} \leq \Delta_\epsilon \leq \min_{n\geq k} \Delta_{\epsilon,\mathcal{C}},$$

then Fig. 7.8a suggests that, for any ϵ, if we use the optimal blocklength, then random codes achieve an age-performance close to the optimal linear code.

Fig. 7.7 and Fig. 7.8b mirror Fig. 7.6 and Fig. 7.8a respectively, but for a larger channel-input alphabet with $q = 25$. We can apply the same analysis as the one we just presented for the case $q = 5$. In this case we can notice the effect of increasing the size of the channel-input alphabet, while keeping k constant. In fact, comparing Fig. 7.6 and Fig. 7.7, we observe a clear convergence of $\Delta_{\epsilon,\mathcal{C}}$ toward the lower bound $\Delta^{lb}_{\epsilon,(n,k)}$. In Fig. 7.7, the approximation $\Delta^{*}_{\epsilon,(n,k)}$ is not as tight as for the case of $q = 5$, for all ϵ and n. Indeed, we can notice that, for $\epsilon = 0.9$, $\Delta^{*}_{\epsilon,(n,k)}$ is worse than $\Delta^{ub}_{\epsilon,\mathcal{C}}$ for $n \leq 20$. For large n, all bounds are tight except for the upper bound $\Delta^{ub}_{\epsilon,\mathcal{C}}$. In fact, in Fig. 7.7, the lower bound $\Delta^{lb}_{\epsilon,(n,k)}$ is the tightest bound on $\Delta_{\epsilon,\mathcal{C}}$. However, the convergence of $\Delta_{\epsilon,\mathcal{C}}$ toward the lower bound $\Delta^{lb}_{\epsilon,(n,k)}$ is best observed in Fig. 7.8b. In this figure, we remark that the performance of the random code with the optimal blocklength is almost optimal. These simulations support our claim that random codes are age-optimal as q grows and the channel-input alphabet becomes large.

7.5 Conclusion

In this chapter, we have studied the optimal achievable average age over an erasure channel in two scenarios: In the first scenario we have considered the source alphabet and channel-input alphabet to be the same. Whereas, in the second scenario, we have assumed the source alphabet to be different than the channel-input alphabet. We have demonstrated that in the first case, we do not need any type of channel coding to achieve the minimal average age, for which we have computed the exact expression. As for the second case, we have used the random coding technique to compute bounds on the optimal achievable age. We have also shown that random codes are age-optimal for large enough channel-input alphabet. Finally, the numerical results have pointed out an interesting observation: Even for a small channel-input alphabet, the performance of random codes is not too far from optimal from an age point of view.

7.6 Appendix: On the Equidistribution Theory

7.6.1 Equidistribution and Weyl's Equidistribution Theorem

In this section[5], for any real number x, we use $[x]$ to denote its fractional part, i.e. $[x] = x - \lfloor x \rfloor$.

Definition 7.9. *A sequence $(u_i)_{i\geq 1} \in [0,1]$ is said to be equidistributed on $[0,1]$ if for any interval $(a,b) \subset [0,1]$ we have*

$$\lim_{N\to\infty} \frac{1}{N} \left|\{1 \leq i \leq N; u_i \in (a,b)\}\right| = b - a, \tag{7.39}$$

where $|A|$ denotes the cardinality of the set A.

Lemma 7.10. *Definition 7.9 implies that we can replace (a,b) with $[a,b)$, $(a,b]$ or $[a,b]$ in (7.39) and the limit will still hold. We will prove this claim for the interval $[a,b]$.*

Proof. For any $\alpha, \epsilon \in (0,1)$,

$$\lim_{N\to\infty} \frac{1}{N} \left|\left\{1 \leq i \leq N; u_i \in \left(\alpha - \frac{\epsilon}{2}, \alpha + \frac{\epsilon}{2}\right)\right\}\right| \leq \epsilon.$$

However,

$$0 \leq \frac{1}{N} \left|\{1 \leq i \leq N; u_i = \alpha\}\right| \leq \frac{1}{N} \left|\left\{1 \leq i \leq N; u_i \in \left(\alpha - \frac{\epsilon}{2}, \alpha + \frac{\epsilon}{2}\right)\right\}\right|.$$

Thus,

$$0 \leq \limsup_{N\to\infty} \left|\{1 \leq i \leq N; u_i = \alpha\}\right| \leq \lim_{N\to\infty} \frac{1}{N} \left|\left\{1 \leq i \leq N; u_i \in \left(\alpha - \frac{\epsilon}{2}, \alpha + \frac{\epsilon}{2}\right)\right\}\right| \leq \epsilon.$$

Since the last inequality is true for any ϵ, this implies

$$\lim_{N\to\infty} \frac{1}{N} \left|\{1 \leq i \leq N; u_i = \alpha\}\right| = 0.$$

[5]The material in this section is based on [7, 69].

For the case of $\alpha = 0$ or $\alpha = 1$, we first notice that $|\{1 \leq i \leq N; u_i \in [0,1]\}| = N$. Hence $\lim_{N\to\infty} \frac{1}{N} |\{1 \leq i \leq N; u_i \in [0,1]\}| = 1$. We also know that

$$\begin{aligned}
1 &= \frac{1}{N} |\{1 \leq i \leq N; u_i \in [0,1]\}| \\
&= \frac{1}{N} |\{1 \leq i \leq N; u_i \in (0,1)\} \cup \{1 \leq i \leq N; u_i = 0\} \cup \{1 \leq i \leq N; u_i = 1\}| \\
&= \frac{1}{N} |\{1 \leq i \leq N; u_i \in (0,1)\}| + \frac{1}{N} |\{1 \leq i \leq N; u_i = 0\}| + \frac{1}{N} |\{1 \leq i \leq N; u_i = 1\}| \\
&\geq \frac{1}{N} |\{1 \leq i \leq N; u_i \in (0,1)\}| .
\end{aligned}$$

This means

$$0 \leq \frac{1}{N} |\{1 \leq i \leq N; u_i = 0\}| + \frac{1}{N} |\{1 \leq i \leq N; u_i = 1\}| \leq 1 - \frac{1}{N} |\{1 \leq i \leq N; u_i \in (0,1)\}| .$$

Since the sequence $(u_i)_{i\geq 1}$ is equidistributed over $[0,1]$ then by (7.39) we have that

$$\lim_{N\to\infty} \frac{1}{N} |\{1 \leq i \leq N; u_i \in (0,1)\}| = 1.$$

Hence, $\lim_{N\to\infty} \frac{1}{N} |\{1 \leq i \leq N; u_i = 0\}| + \frac{1}{N} |\{1 \leq i \leq N; u_i = 1\}| = 0$. This implies that $\lim_{N\to\infty} \frac{1}{N} |\{1 \leq i \leq N; u_i = 0\}| = 0$ and $\lim_{N\to\infty} \frac{1}{N} |\{1 \leq i \leq N; u_i = 1\}| = 0$.

Finally, for any $a, b \in [0,1]$,

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N} |\{1 \leq i \leq N; u_i \in [a,b]\}| \\
&= \lim_{N\to\infty} \frac{1}{N} |\{1 \leq i \leq N; u_i \in (a,b)\} \cup \{1 \leq i \leq N; u_i = a\} \cup \{1 \leq i \leq N; u_i = b\}| \\
&= \lim_{N\to\infty} \frac{1}{N} |\{1 \leq i \leq N; u_i \in (a,b)\}| + \frac{1}{N} |\{1 \leq i \leq N; u_i = a\}| + \frac{1}{N} |\{1 \leq i \leq N; u_i = b\}| \\
&= \lim_{N\to\infty} \frac{1}{N} |\{1 \leq i \leq N; u_i \in (a,b)\}| .
\end{aligned}$$

□

Proposition 7.2. *If $(u_i)_{i\geq 1}$ is an equidistributed sequence on $[0,1]$ then the set $\{(u_i)_{i\geq 1}\}$ is dense in $[0,1]$.*

Proof. Fix $\alpha \in (0,1)$ and $\epsilon \in (0,1)$. Since $(u_i)_{i\geq 1}$ are equidistributed over $[0,1]$ then by (7.39)

$$\lim_{N\to\infty} \left| \left\{ 1 \leq i \leq N; u_i \in \left(\alpha - \frac{\epsilon}{2}, \alpha + \frac{\epsilon}{2}\right) \right\} \right| = \epsilon > 0.$$

This means that $\forall \delta > 0, \exists N_0(\delta) \in \mathbb{N}$ such that $\forall N \geq N_0(\delta)$

$$\left| \frac{1}{N} \left| \left\{ 1 \leq i \leq N; u_i \in \left(\alpha - \frac{\epsilon}{2}, \alpha + \frac{\epsilon}{2}\right) \right\} \right| - \epsilon \right| < \delta.$$

Hence, $\forall \delta > 0$, $\exists j_\delta \geq N_0(\delta)$ such that

$$u_{j_\delta} \in \left(\alpha - \frac{\epsilon}{2}, \alpha + \frac{\epsilon}{2}\right) \qquad \text{and} \qquad |u_{j_\delta} - \alpha| < \epsilon.$$

Given that the choice of ϵ is arbitrary, then there exists a subsequence (u_{j_δ}) of $(u_i)_{i\geq 1}$ that converges to α.

If $\alpha = 0$ we use Lemma 7.10. Thus,

$$\lim_{N\to\infty} |\{1 \leq i \leq N; u_i \in (0,\epsilon)\}| = \lim_{N\to\infty} |\{1 \leq i \leq N; u_i \in [0,\epsilon)\}| = \epsilon > 0.$$

Hence, $\forall \delta > 0$, $\exists j_\delta \geq N_0(\delta)$ such that

$$u_{j_\delta} \in [0,\epsilon) \qquad \text{and} \qquad |u_{j_\delta}| < \epsilon.$$

Given that the choice of ϵ is arbitrary, then there exists a subsequence (u_{j_δ}) of $(u_i)_{i\geq 1}$ that converges to 0. A similar argument can be applied for $\alpha = 1$. This proves our lemma. □

Theorem 7.7. *Let $(u_i)_{i\geq 1}$ be a sequence of real numbers and denote by $[u_i] = u_i - \lfloor u_i \rfloor$ the fractional part of u_i. Then the following are equivalent:*

1. *The sequence $([u_i])_{i\geq 1}$ is equidistributed on $[0,1]$.*
2. *For any $k \in \mathbb{N}^*$,*
$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} e^{j2\pi k u_i} = 0, \tag{7.40}$$
where $j^2 = -1$.
3. *For any Riemann-integrable function $f : [0,1] \to \mathbb{C}$ we have*
$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} f([u_i]) = \int_0^1 f(x)\mathrm{d}x. \tag{7.41}$$

The proof of Theorem 7.7 is outside the scope of this text but we encourage the reader to check [7] for the full proof. An important application of this theorem is given next.

Corollary 7.3. *Let α be an irrational number. Then the sequence $([n\alpha])_{n\geq 0}$ is equidistributed over $[0,1]$.*

Proof. We use the second criterion of Theorem 7.7 with $(u_i)_{i\geq 1} = ((i-1)\theta)_{i\geq 1}$. Hence,

$$\begin{aligned}
\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} e^{j2\pi k u_i} &= \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} e^{j2\pi k(i-1)\alpha} \\
&= \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} e^{j2\pi kn\alpha} \\
&= \lim_{N\to\infty} \frac{1}{N} \left(\frac{1 - e^{j2\pi k\alpha N}}{1 - e^{j2\pi k\alpha}} \right) \\
&= 0,
\end{aligned}$$

where the last inequality is justified by the fact that $e^{j2\pi k\alpha} \neq 1$ for any $k \in \mathbb{N}^*$ since θ is irrational and that $e^{j2\pi k\alpha N}$ has a modulus of 1 for any k and N. □

7.6.2 Proof of Lemma 7.1

Let α be an irrational number. From Corollary 7.3 we know that the sequence $([i\alpha])_{i\geq 1}$ is equidistributed over $[0,1]$. Therefore, using the third criterion from Theorem 7.7 with the Riemann-integrable function f being the identity function, i.e.

$$\begin{aligned} f : [0,1] &\to [0,1] \\ x &\mapsto x, \end{aligned}$$

we get

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} [i\alpha] = \int_0^1 x \mathrm{d}x = \frac{1}{2}.$$

7.6.3 Proof of Lemma 7.2

Before presenting the proof of Lemma 7.2 proper, we need the following lemma.

Lemma 7.11. *Let $\alpha \in \mathbb{R}\setminus\mathbb{Q}$ be a rational positive number. Assume also that $\alpha = \frac{m}{l}$, with $m, l \in \mathbb{N}$, $l \neq 0$ and $\gcd(m,l) = 1$. Then, for $x = 0, 1, \cdots, l-1$,*

$$\lim_{N\to\infty} \frac{1}{N} \left| \left\{ 0 \leq i \leq N-1; [i\alpha] = \frac{x}{l} \right\} \right| = \frac{1}{l}, \tag{7.42}$$

with $[i\alpha]$ being the fractional part of $i\alpha$.
This lemma says that the fractional part of $i\alpha$, for any integer i, can only take values in $\{0, \frac{1}{l}, \cdots, \frac{l-1}{l}\}$ and in a uniform manner.

Proof. We have that $\alpha = \frac{m}{l}$ with $m, l \in \mathbb{N}$ and $\gcd(m,l) = 1$. First notice that for any $i \in \mathbb{N}$, we can write

$$i\alpha = \frac{im}{l} = \frac{pl + r}{l} = p + \frac{r}{l},$$

where $p, r \in \mathbb{N}$ and $0 \leq r \leq (l-1)$. Thus,

$$[i\alpha] = \frac{im \ [\mathrm{mod}\ l]}{l}.$$

Hence the quantity of interest is $im \ [\mathrm{mod}\ l]$ for any $i \in \mathbb{N}$. However, $\forall i \in \mathbb{N}$, $im \ [\mathrm{mod}\ l] \in \{0, 1, \cdots, l-1\}$. This means that there should be at least two integers i_1 and i_2 such that $i_1 \neq i_2$ and $i_1 m \equiv i_2 m \ [\mathrm{mod}\ l]$. Without loss of generality, assume that $i_2 > i_1$. This means that there exists at least one integer $n = i_2 - i_1 \neq 0$ such that $nm \equiv 0 \ [\mathrm{mod}\ l]$. Let $\hat{n}$ be the smallest integer different than 0 such that $\hat{n}m \equiv 0 \ [\mathrm{mod}\ l]$, i.e. $\hat{n} = \min\{n; n \neq 0, nm \equiv 0 \ [\mathrm{mod}\ l]\}$.

Fact 7.1. *In the set $\{0, m, 2m, \cdots, (\hat{n}-1)m\}$ there cannot be any two elements that are equal modulo l. This means that for any $k_1 \neq k_2$ and $k_1, k_2 \in \{0, 1, \cdots, \hat{n}-1\}$,*

$$k_1 m \not\equiv k_2 m \ [\mathrm{mod}\ l].$$

Proof of Fact 7.1. To prove this claim let's assume that there exist two integers $k_1 \neq k_2$, $k_1, k_2 \in \{0, 1, \cdots, \hat{n}-1\}$ and $k_1 m \equiv k_2 m \ [\text{mod } l]$. Without loss of generality, take $k_2 > k_1$. Hence, $\hat{k} = k_2 - k_1 \neq 0$ is such that $\hat{k}m \equiv 0 \text{ mod } l$. Moreover, $\hat{k} < \hat{n}$ which contradicts the minimality of $\hat{n}$. Therefore, no two elements of $\{0, m, 2m, \cdots, (\hat{n}-1)m\}$ are equal modulo l. □

Fact 7.2. *If* $\hat{n} = \min\{n; n \neq 0, nm \equiv 0 \ [\text{mod } l]\}$, *then* $\hat{n} = l$.

Proof of Fact 7.2. From the definition of $\hat{n}$ and the fact that $lm \equiv 0 \ [\text{mod } l]$, we know using Fact 7.1 that $\hat{n} \leq l$. Using Bézout's identity and given that $\gcd(m, l) = 1$, we know that there exist two integers n_1 and n_2 such that $n_1 l + n_2 m = 1$. Hence $n_2 m \equiv 1 \ [\text{mod } l]$. Given that $n_2 m \equiv 1 \ [\text{mod } l]$, then $jn_2 m \equiv j \ [\text{mod } l]$, $\forall j \in \{0, 1, \cdots, l-1\}$. This means that $\hat{n} = l$ and

$$\{0, m \ [\text{mod } l], 2m \ [\text{mod } l], \cdots, (\hat{n}-1)m \ [\text{mod } l]\} = \{0, 1, \cdots, l-1\}.$$

This can be explained using the following argument: For any integer j, $jn_2 = x_j \hat{n} + y_j$ with $x_j \in \mathbb{Z}$ and $0 \leq y_j \leq \hat{n} - 1$. Hence, $jn_2 m \equiv y_j m \ [\text{mod } l] \equiv j \ [\text{mod } l]$ since $x_j \hat{n} m \equiv 0 \ [\text{mod } l]$. Given that the set $\{y_j m \ [\text{mod } l]; 0 \leq j \leq l-1\}$ has l distinct elements, this means that $\hat{n} \geq l$. But $\hat{n} \leq l$. Therefore, $\hat{n} = l$. □

Fact 7.2 shows that the sequence $(im \ [\text{mod } l])_{i \geq 0}$ is a periodic sequence with period l. Moreover, in each period, we visit each element of the set $\{0, 1, \cdots, l-1\}$ exactly once. This means that over a discrete interval of length N, the number of time an integer $x \in \{0, 1, \cdots, l-1\}$ is visited can be bounded as follows:

$$\frac{N-l}{l} \leq \left| \left\{ 0 \leq i \leq N-1; [i\alpha] = \frac{im \ [\text{mod } l]}{l} = \frac{x}{l} \right\} \right| \leq \frac{N+l}{l}$$
$$\frac{N}{l} - 1 \leq \left| \left\{ 0 \leq i \leq N-1; [i\alpha] = \frac{im \ [\text{mod } l]}{l} = \frac{x}{l} \right\} \right| \leq \frac{N}{l} + 1.$$

Therefore,

$$\frac{1}{l} - \frac{1}{N} \leq \frac{1}{N} \left| \left\{ 0 \leq i \leq N-1; [i\alpha] = \frac{im \ [\text{mod } l]}{l} = \frac{x}{l} \right\} \right| \leq \frac{1}{l} + \frac{1}{N}.$$

Hence,

$$\lim_{N \to \infty} \frac{1}{N} \left| \left\{ 0 \leq i \leq N-1; [i\alpha] = \frac{x}{l} \right\} \right| = \frac{1}{l}.$$

This proves our lemma.

□

Now we can give the proof for Lemma 7.2.

Proof of Lemma 7.2. First notice that (7.42) can be written as

$$\lim_{N \to \infty} \frac{1}{N} \left| \left\{ 0 \leq i \leq N-1; [i\alpha] = \frac{x}{l} \right\} \right| = \lim_{N \to \infty} \frac{1}{N} \sum_{i=0}^{N-1} \mathbb{1}_{\{[i\alpha] = \frac{x}{l}\}} = \frac{1}{l}, \tag{7.43}$$

with $\mathbb{1}_{\{.\}}$ is the indicator function. Hence, using Lemma 7.11,

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\sum_{i=0}^{N-1}[i\alpha] &= \lim_{N\to\infty}\frac{1}{N}\sum_{i=0}^{N-1}\sum_{x=0}^{l-1}\frac{x}{l}\mathbb{1}_{\{[i\alpha]=\frac{x}{l}\}}\\ &\overset{(a)}{=}\sum_{x=0}^{l-1}\frac{x}{l}\left[\lim_{N\to\infty}\frac{1}{N}\sum_{i=0}^{N-1}\mathbb{1}_{\{[i\alpha]=\frac{x}{l}\}}\right]\\ &\overset{(b)}{=}\sum_{x=0}^{l-1}\frac{x}{l^2}\\ &=\frac{l-1}{2l},\end{aligned}$$

where the exchange of the sum and limit in (a) is justified by the fact that the sum over x is finite. Moreover, the equality in (b) is due to (7.43). □

Conclusion and Further Directions

8

8.1 Conclusion

In this thesis, we have used the age-of-information metric in order to assess the performance of different communication systems. The two main computational approaches of this metric have been presented in Chapter 2 and have constituted the framework we have used, in the various chapters of this work, to derive two age metrics: the average age and/or the average peak-age. The communication systems that we have considered in this dissertation are divided into two categories: systems with a noiseless channel and systems with a noisy channel.

In the first part of the thesis, we have focused on systems with a noiseless channel. This means that any transmitted packet is received at the monitor with probability 1, unless the transmitter decides to preempt it. For such a model, we use a queue-theoretic approach to analyze the system and compute the average age and the average peak-age. Indeed, in this part, we assume the packets are generated according to a Poisson process with rate λ and transmitted to the monitor through a network whose service-time depends on the studied transmission scheme. In Chapter 3, we have assumed the network has a gamma distributed service-time, and we have considered two transmission policies: LCFS with preemption and with-preemption-in-waiting. In this chapter, we have shown that the LCFS with-preemption-in-waiting policy achieves in general an average age lower than the LCFS with-preemption scheme. Using our results for gamma distributed service-time, we derive exact expressions of the average age and average peak-age under deterministic service-time. We notice that for the LCFS with-preemption scheme, deterministic service-time leads to a higher average age, compared to the one achieved when assuming any regular gamma distribution. Whereas, for the LCFS with-preemption-in-waiting policy, the average age attained by a system with deterministic service-time is lower than the one attained by a system with gamma distributed service-time.

In Chapter 4, we have generalized a section of the results of Chapter 3 in two

directions: (*i*) We consider that the sender comprises multiple sources instead of only one packet source, and (*ii*) we assume a general distribution for the service-time. The transmission policy that we have studied is the LCFS with preemption or M/G/1/1 with preemption (in queue-theoretic terms). This means that each source i generates updates according to a Poisson process with rate λ_i, but all packets are transmitted according to the same service-time distribution. In order to compute the average age and the average peak-age of such a model, we introduce the detour-flow graph method as a tool that could be used to solve age-of-information problems. We have demonstrated that the average age relative to each source i depends only on λ_i and the Laplace transform of the service-time distribution. We have considered a new age metric which is the sum of all individual average ages. We have shown that in order to minimize this metric, all sources should generate packets at the same rate.

In the analysis presented in Chapter 4, we have looked at the different sources as being interchangeable with all of them enjoying the same priority level. However, in practice, this assumption is not always accurate. When multiple data streams share the same transmitter, some streams have contents more important than others thus are assigned a higher transmission priority. In Chapter 5, we consider a sender with two sources with different priorities. A source with a high priority, whose packets are given precedence over those of another source, tagged as low priority. We have assumed an M/G/1/1 transmission scheme for the high-priority source and have studied two different transmission policies for the low-priority source: an FCFS with exponential service time and an M/G/1/1 with preemption. For the FCFS model, we have presented the stability condition needed for the queue of the low-priority stream to stay stable and to give bounds on the average age. Using the detour-flow graph method of Chapter 4 and assuming M/G/1/1 with preemption, we have computed the average age and average peak-age relative to the low-priority source. It has also been shown that the introduction of a higher priority stream into an age-optimal system renders it suboptimal. Indeed, it is shown in [6] that under exponential service-time, the LCFS with preemption policy is optimal. Nevertheless, in Chapter 5 we observe that, under exponential service-time, the average age of the low-priority stream, when we consider an FCFS model, is lower than the average age of this stream when we assume a preemption model.

In the second part of this thesis, we have considered a faulty channel. This means that some transmitted packets could be erased and never delivered to the receiver. We have modeled this concept by adopting an erasure channel as the transmission medium. In Chapter 6, we have applied the age-of-information metrics to assess the performance of two error-correcting protocols that are used in real-life communication systems (such as 5G and Wimax): HARQ with infinite incremental redundancy (IIR-HARQ) and HARQ with fixed redundancy (HARQ-FR). For each of these protocols, we have considered two transmission schemes: M/G/1/1 with blocking and M/G/1/1 with preemption. We have computed the average age for each one of the four combinations and have observed the following: First, for any of the two error correcting protocols, the M/G/1/1 with blocking scheme achieves an average age lower than that of the M/G/1/1 with preemption. Second, for any of the two transmission schemes, IIR-HARQ exhibits a better performance, from an age point of view, than FR-HARQ.

Up till here, we have assumed random update-generation mechanisms. In particular, we have considered, in the previous chapters, that packets are generated according to a Poisson process. In Chapter 7, we have adopted a different approach: A single source generates updates periodically in a deterministic manner and the transmitter also has periodic access to the channel. As in Chapter 6, we have transmitted over an erasure channel, but this time with no feedback. Given such a model, we have answered the following question: What is the best average age that could be achieved on an erasure channel with no feedback? We have divided this problem into two scenarios: In the first scenario, the source alphabet and the channel-input alphabet are the same, whereas in the second scenario they are different. In the first scenario, we have proved that the optimal average age is achieved without any channel coding and we have derived a closed-form expression of the average age. For the second scenario, finding the optimal coding scheme constitutes a more difficult challenge. To solve this problem, we have called on an information-theoretic tool: A random coding argument is used to give bounds on the optimal achievable age. We have also shown that random linear codes achieve the optimal age for large enough channel-input alphabet. This means, that random codes are not just tools for existential proof but also explicit age-optimal error correcting codes.

8.2 Further Directions

In the Absence of Noise

In Chapter 4, we have generalized only the results related to the LCFS with preemption scheme. However, we have seen in Chapter 3, the LCFS without preemption or M/M/1/2* scheme exhibits a better performance from an age point of view. Therefore, it would be interesting to compute the average age of a multi-stream M/G/1/2* system and compare it to the average age of a multi-stream M/G/1/1 preemptive scheme, for general service time. Yates et al. solve this problem in [75] when the service time is exponentially distributed. However, we could then answer the question: For what types of service times is preemption in service preferable over preemption in waiting? A further generalization of the result in Chapter 4 would be to consider a multi-stream system with a G/G/1/1 with preemption scheme.

In Chapter 5, we have seen that, for the low-priority stream, adopting an FCFS policy is more beneficial than an M/M/1/1 with preemption scheme. This leads us to think that an M/M/1/2* strategy would outperform both previous policies because it appears to have all the advantages of an FCFS policy and reduces the waiting time of the packets since only the freshest update is stored. Therefore, we could compute the average age of the low priority stream under a M/G/1/2* strategy and compare its performance to the FCFS and M/G/1/1 with preemption policies, when we assume an exponential service time. Moreover, in Chapter 5, we have given bounds only on the average age of the low priority stream. It would be interesting to derive a closed-form expression of this average age, at least when we assume exponential service time.

In the Presence of Noise

In Chapter 7 we have shown that random codes are age optimal when the channel-input alphabet is large enough. However, for a small channel-input alphabet, random codes give only an upper bound on the optimal achievable age and the search for the optimal channel code is still open. Indeed, for low erasure probability, we believe that we will achieve an average age lower than the one achieved with random codes, if we use an (n, k)-linear code that transmits first the systematic part of the codeword (this means the part that corresponds to the k information symbols) and then randomly chooses the $n - k$ remaining symbols of the codeword. The intuition behind this is as follows: For a low erasure probability, we have a high probability of correctly receiving any transmitted symbol. Therefore, by sending the systematic part first, we have a high probability of correctly decoding the transmitted codeword after just k channel uses. This would lead to a lower average age, compared to a full random code that might need more than k channel uses to send the first k linearly independent coded symbols (as it might generate linearly dependent symbols among the first k coded symbols).

The presence of noise in the channel opens multiple research paths on the age of information. In this thesis, we have considered only erasure channels because the concept of age is easily defined in this case: a packet is either received or erased. However, we could consider various types of channels such as binary symmetric channels (BSC) or simply a general channel modeled by a stochastic matrix. In this case, a new definition of the average age needs to be developed: This new definition should assign some age penalty to the packet decoded incorrectly, as they will not convey the correct information about the source. One approach would be to keep the definition of the average age as presented in Chapter 1, but we add a second metric that measures the confidence (or the precision) of the decoded packets. Therefore, a communication system would be characterized by a couple of two metrics: the average age that indicates how fresh the receiver's information about the observed process is, and a confidence measure that reflects the level of decoding error. We believe that these two metrics have a negative correlation. This means that any scheme that aims at reducing the average age would lead to higher decoding error rates, and that any coding scheme that reduces the decoding error would increase the average age.

Finally, we can think of the average age as a measure of the distortion of the information in time (in contrast to the distortion in content that is usually studied in information theory). This view on the age raises an interesting rate-distortion kind of problem: Given two random processes $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$, what is the minimum needed rate between these two processes to achieve an average age less than or equal to a given threshold. Formally, let X^n and Y^n be two processes, with the X_i's drawn from alphabet $\mathcal{X}$ and the Y_i's drawn from an alphabet $\mathcal{Y}$. We assume we are given the distribution of X^n. For now, we will consider that the X_i are i.i.d distributed according to p_X. We define a distortion measure between X^n and Y^n, denoted by

Definition 8.1.

$$Age(X^n, Y^n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(i - J_i),$$

where

$$J_i = \max_{\substack{j \leq i \ \text{s.t.} \\ \mathbb{P}_{Y^i|X_j}(y^i|x), \forall x \in \mathcal{X} \setminus \{X_j\}}} j.$$

This means that J_i is the index j of the highest correctly retrieved X_j.

Let $\mathcal{W}_{Y^n|X^n}(y^n|x^n)$ be the channel that outputs y^n given a certain sequence x^n. Based on Definition 8.1, we can see that the age is function of the channel $\mathcal{W}$.

Definition 8.2. *We define the rate-age function $R(D)$ to be*

$$R(D) = \min_{\substack{\mathcal{Y},\ \mathcal{W}_{Y^n|X^n} \ \text{s.t.} \\ \mathcal{W}_{Y^n|X^n} \text{ causal}, \\ Age(\mathcal{W}_{Y^n|X^n}) \leq D}} \frac{1}{n} I(X^n; Y^n).$$

Our aim would be to compute and characterize this rate-age function $R(D)$.

Bibliography

[1] A. Arafa and S. Ulukus, "Age-minimal transmission in energy harvesting two-hop networks," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017, pp. 1–6.

[2] B. T. Bacinoglu, E. T. Ceran, and E. Uysal-Biyikoglu, "Age of information under energy replenishment constraints," in *Proc. Info. Theory and Appl. (ITA) Workshop*, Feb. 2015, la Jolla, CA.

[3] B. T. Bacinoglu, Y. Sun, E. Uysal-Bivikoglu, and V. Mutlu, "Achieving the age-energy tradeoff with a finite-battery energy harvesting source," in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 876–880.

[4] B. T. Bacinoglu and E. Uysal-Biyikoglu, "Scheduling status updates to minimize age of information with an energy harvesting sensor," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 1122–1126.

[5] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Optimizing data freshness, throughput, and delay in multi-server information-update systems," in *Proc. IEEE Int'l. Symp. Info. Theory*, 2016, pp. 2569–2574.

[6] ——, "Age-optimal information updates in multihop networks," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 576–580.

[7] K. Chandrasekharan, *Introduction to analytic number theory*, ser. Grundlehren der mathematischen Wissenschaften. Spring-Verlag, 1968. [Online]. Available: https://books.google.ch/books?id=h9ZQAAAAMAAJ

[8] K. Chen and L. Huang, "Age-of-information in the presence of error," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 2579–2583.

[9] M. Costa, M. Codreanu, and A. Ephremides, "Age of information with packet management," in *Proc. IEEE Int'l. Symp. Info. Theory*, June 2014, pp. 1583–1587.

[10] ——, "On the age of information in status update systems with packet management," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 1897–1910, April 2016.

[11] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., 2006.

[12] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011. [Online]. Available: https://books.google.ch/books?id=2gsLkQlb8JAC

[13] J. Daigle, *Queueing Theory with Applications to Packet Telecommunication*, ser. Queueing Theory with Applications to Packet Telecommunication. Springer, 2005. [Online]. Available: https://books.google.com.lb/books?id=EkSsgr-d8ZoC

[14] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal-Biyikoglu, "Delay and peak-age violation probability in short-packet transmissions," in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 2471–2475.

[15] S. Feng and J. Yang, "Optimal status updating for an energy harvesting sensor with a noisy channel," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, April 2018, pp. 348–353.

[16] R. G. Gallager, *Discrete Stochastic Processes*, 2nd ed. Kluwer Academic Publishers, Feb. 1996.

[17] Q. He, D. Yuan, and A. Ephremides, "On optimal link scheduling with min-max peak age of information in wireless systems," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–7.

[18] ——, "Optimal link scheduling for age minimization in wireless systems," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 5381–5394, July 2018.

[19] T. He, S. Krishnamurthy, J. A. Stankovic, T. Abdelzaher, L. Luo, R. Stoleru, T. Yan, L. Gu, J. Hui, and B. Krogh, "Energy-efficient surveillance system using wireless sensor networks," in *Proceedings of the 2nd international conference on Mobile systems, applications, and services*. ACM, 2004, pp. 270–283.

[20] M. Heindlmaier and E. Soljanin, "Isn't hybrid ARQ sufficient?" in *Communication, Control, and Computing (Allerton), 52nd Annual Allerton Conference on*. IEEE, 2014, pp. 563–568.

[21] J. P. Hespanha, "Modelling and analysis of stochastic hybrid systems," *IEE Proceedings - Control Theory and Applications*, vol. 153, no. 5, pp. 520–535, Sept 2006.

[22] L. Huang and E. Modiano, "Optimizing age-of-information in a multi-class queueing system," in *2015 IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 1681–1685.

[23] Y. Inoue, H. Masuyama, T. Takine, and T. Tanaka, "The stationary distribution of the age of information in fcfs single-server queues," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 571–575.

[24] I. Kadota, A. Sinha, and E. Modiano, "Optimizing age of information in wireless networks with throughput constraints," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, April 2018, pp. 1844–1852.

[25] I. Kadota, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Minimizing the age of information in broadcast wireless networks," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2016, pp. 844–851.

[26] C. Kam, S. Kompella, and A. Ephremides, "Age of information under random updates," in *Proc. IEEE Int'l. Symp. Info. Theory*, 2013, pp. 66–70.

[27] ——, "Effect of message transmission diversity on status age," in *Proc. IEEE Int'l. Symp. Info. Theory*, June 2014, pp. 2411–2415.

[28] C. Kam, S. Kompella, G. D. Nguyen, and A. Ephremides, "Effect of message transmission path diversity on status age," *IEEE Transactions on Information Theory*, vol. 62, no. 3, pp. 1360–1374, March 2016.

[29] C. Kam, S. Kompella, G. D. Nguyen, J. E. Wieselthier, and A. Ephremides, "Age of information with a packet deadline," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 2564–2568.

[30] ——, "Controlling the age of information: Buffer size, deadline, and packet replacement," in *MILCOM 2016 - 2016 IEEE Military Communications Conference*, Nov 2016, pp. 301–306.

[31] ——, "On the age of information with packet deadlines," *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6419–6428, Sept 2018.

[32] ——, "Towards an effective age of information: Remote estimation of a markov source," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, April 2018, pp. 367–372.

[33] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, Salt Lake City, Utah, USA, 2011.

[34] S. Kaul, R. D. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. INFOCOM*, 2012.

[35] ——, "Status updates through queues," in *Conf. on Information Sciences and Systems (CISS)*, Mar. 2012.

[36] S. K. Kaul and R. D. Yates, "Age of information: Updates with priority," in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 2644–2648.

[37] S. K. Kaul, R. D. Yates, and M. Gruteser, "On piggybacking in vehicular networks," in *IEEE Global Telecommunications Conference, GLOBECOM 2011*, Dec. 2011.

[38] S. Kaul, R. Yates, and M. Gruteser, "Status updates through queues," in *Information Sciences and Systems (CISS), 2012 46th Annual Conference on*, March 2012, pp. 1–6.

[39] S. Kim, S. H. Son, J. A. Stankovic, S. Li, and Y. Choi, "SAFE: a data dissemination protocol for periodic updates in sensor networks," in *23rd International Conference on Distributed Computing Systems Workshops, 2003. Proceedings.* IEEE, May 2003, pp. 228–234.

[40] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, "Age and value of information: Non-linear age case," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 326–330.

[41] A. Kosta, N. Pappas, and V. Angelakis, "Age of information: A new concept, metric, and tool," *Foundations and Trends in Networking*, vol. 12, no. 3, pp. 162–259, 2017. [Online]. Available: http://dx.doi.org/10.1561/1300000060

[42] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson, "Wireless sensor networks for habitat monitoring," in *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications.* ACM, 2002, pp. 88–97.

[43] P. Mayekar, P. Parag, and H. Tyagi, "Optimal lossless source codes for timely updates," in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 1246–1250.

[44] E. Najm and R. Nasser, "Age of information: The gamma awakening," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 2574–2578.

[45] E. Najm, R. Nasser, and E. Telatar, "Content based status updates," in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 2266–2270.

[46] E. Najm and E. Telatar, "Status updates in a multi-stream M/G/1/1 preemptive queue," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, April 2018, pp. 124–129.

[47] E. Najm, R. Yates, and E. Soljanin, "Status updates through M/G/1/1 queues with HARQ," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 131–135.

[48] E. Najm, R. D. Yates, and E. Soljanin, "Status updates through M/G/1/1 queues with HARQ," *ArXiv e-prints*, Apr. 2017.

[49] E. Najm and R. Nasser, "Age of information: The gamma awakening," *CoRR*, vol. abs/1604.01286, 2016. [Online]. Available: http://arxiv.org/abs/1604.01286

[50] E. Najm, R. Nasser, and E. Telatar, "Content based status updates," *CoRR*, vol. abs/1801.04067, 2018. [Online]. Available: http://arxiv.org/abs/1801.04067

[51] P. Papadimitratos, A. La Fortelle, K. Evenssen, R. Brignolo, and S. Cosenza, "Vehicular communication systems: Enabling technologies, applications, and future outlook on intelligent transportation," *IEEE Communications Magazine*, vol. 47, no. 11, pp. 84–95, Nov. 2009.

[52] N. Pappas, J. Gunnarsson, L. Kratz, M. Kountouris, and V. Angelakis, "Age of information of multiple sources with queue management," in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 5935–5940.

[53] P. Parag, A. Taghavi, and J. Chamberland, "On real-time status updates over symbol erasure channels," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, March 2017, pp. 1–6.

[54] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[55] T. Richardson and R. Urbanke, *Modern Coding Theory*. New York, NY, USA: Cambridge University Press, 2008.

[56] B. Rimoldi, *Principles of Digital Communication: A Top-Down Approach*. Cambridge University Press, 2016.

[57] P. E. Ross, "Managing care through the air [remote health monitoring]," *IEEE Spectrum*, vol. 41, no. 12, pp. 26–31, Dec. 2004.

[58] S. M. Ross, *Stochastic Processes (Wiley Series in Probability and Statistics)*, 2nd ed. Wiley, Feb. 1995.

[59] H. Sac, T. Bacinoglu, E. Uysal-Biyikoglu, and G. Durisi, "Age-optimal channel coding blocklength for an M/G/1 queue with HARQ," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2018, pp. 1–5.

[60] C. Schurgers, V. Tsiatsis, S. Ganeriwal, and M. Srivastava, "Optimizing sensor networks in the energy-latency-density design space," *IEEE Transactions on Mobile Computing*, vol. 1, no. 1, pp. 70–80, Mar. 2002.

[61] M. Shaked and J. G. Shanthikumar, *Stochastic Orders*. Springer Science and Business Media, 2007.

[62] S. Shamai, I. E. Telatar, and S. Verdu, "Fountain capacity," *IEEE Transactions on Information Theory*, vol. 53, no. 11, pp. 4372–4376, Nov 2007.

[63] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x

[64] A. Soysal and S. Ulukus, "Age of information in G/G/1/1 systems," *CoRR*, vol. abs/1805.12586, 2018. [Online]. Available: http://arxiv.org/abs/1805.12586

[65] Y. Sun, Y. Polyanskiy, and E. Uysal-Biyikoglu, "Remote estimation of the wiener process over a channel with random delay," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 321–325.

[66] Y. Sun, E. Uysal-Biyikoglu, R. D. Yates, C. E. Koksal, and N. B. Shroff, "Update or wait: How to keep your data fresh," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, April 2016, pp. 1–9.

[67] ——, "Update or wait: How to keep your data fresh," *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 7492–7508, Nov 2017.

[68] R. Talak, S. Karaman, and E. Modiano, "Can Determinacy Minimize Age of Information?" *ArXiv e-prints*, Oct. 2018.

[69] H. Weyl, "Über die gleichverteilung von zahlen mod. eins," *Mathematische Annalen*, vol. 77, no. 3, pp. 313–352, Sep 1916. [Online]. Available: https://doi.org/10.1007/BF01475864

[70] X. Wu, J. Yang, and J. Wu, "Optimal status update for age of information minimization with an energy harvesting source," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 1, pp. 193–204, March 2018.

[71] R. D. Yates, "Lazy is timely: Status updates by an energy harvesting source," in *Proc. IEEE Int'l. Symp. Info. Theory*, 2015.

[72] ——, "Age of information in a network of preemptive servers," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, April 2018, pp. 118–123.

[73] ——, "Status updates through networks of parallel servers," in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 2281–2285.

[74] R. D. Yates and S. Kaul, "Real-time status updating: Multiple sources," in *Proc. IEEE Int'l. Symp. Info. Theory*, Jul. 2012.

[75] R. D. Yates and S. K. Kaul, "The age of information: Real-time status updating by multiple sources," *CoRR*, vol. abs/1608.08622, 2016. [Online]. Available: http://arxiv.org/abs/1608.08622

[76] R. D. Yates, E. Najm, E. Soljanin, and J. Zhong, "Timely updates over an erasure channel," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 316–320.

[77] ——, "Timely updates over an erasure channel," *CoRR*, vol. abs/1704.04155, 2017. [Online]. Available: http://arxiv.org/abs/1704.04155

[78] J. Zhong and R. D. Yates, "Timeliness in lossless block coding," in *2016 Data Compression Conference (DCC)*, March 2016, pp. 339–348.

[79] J. Zhong, R. D. Yates, and E. Soljanin, "Backlog-adaptive compression: Age of information," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 566–570.

[80] ——, "Two freshness metrics for local cache refresh," in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 1924–1928.

Curriculum Vitae

Elie Najm

Rue de l'Industrie 9
1020 Renens
Switzerland
Tel: +41786623184
eliegnajm@gmail.com

EDUCATION

PhD Student in Information Theory Lab, EPFL

2014-Present

- *Research area*: Information theory, Age of information, Queuing theory, Probability theory.
- *Supervisor*: Prof. Emre Telatar.
- Head of Teaching Assistants in 5 different courses with responsibilities including: preparing exams, preparing exercise sessions and giving lectures.

Fall 2016-2017

- **One-semester exchange at Rutgers University, NJ, USA, under Prof. Emina Soljanin.**

Masters in Communication Systems, EPFL

2012-2014

- Graduated top 3 of the class, GPA: 5.66/6.0.
- *Specialization*: Wireless Communications, Signal processing.

Bachelor of Computer and Communication Engineering, AUB

2008-2012

- Graduated with High Distinction, top 3 of the class, GPA: 4.0/4.0.
- Completed a minor in Mathematics.

Collège Notre Dame de Jamhour, Lebanon – High School

2005-2008

- Graduated with High Distinction, top of the class.

TEACHING EXPERIENCE

Information Theory and Coding, EPFL

Fall 2017-2018

- Prepared the exercises for the course.
- Supervised the exercise sessions.
- Corrected the exams.

Principles of Digital Communications, EPFL

Spring 2016-2017 and Spring 2017-2018

- Prepared the exercises for the course.
- Supervised the exercise sessions.
- Corrected the exams.

Statistical Signal Processing through Application, EPFL

Spring 2015-2016

- Prepared and gave 8h of lectures.
- Prepared the exercises for the course.
- Supervised the exercise sessions.

Information, Computation, Communication (ICC), EPFL

Fall 2015-2016

- Supervised the exercise sessions.

Linear Algebra I, EPFL

Fall 2014-2015

- Supervised the exercise sessions.

ACADEMIC PROJECTS

Face Recognition on the AT&T Database, EPFL

Spring 2016-2017

- Part of the course *Fundamentals in Statistical Pattern Recognition*.
- Goal: perform face detection given a limited database of pictures corresponding to different persons.
- Implemented in Python three different solutions that use PCA (principle component analysis) as a background model: SVM (Support Vector Machine), NN (Neural Network) and LR (Logistic Regression).
- Achieved a classification error as low as 2.67% using SVM.

Object Recognition on the small NORB 3D Dataset, EPFL

Spring 2012-2013

- Part of the course *Pattern Recognition and Machine Learning.*
- Goal: perform object detection given a limited database of pictures corresponding to 5 different categories.
- Implemented in MATLAB three solutions based on NN (Neural Network), linear regression and logistic regression.
- Achieved an error rate of 18% using Neural Networks.

Proving Strong Resolvability for the BEC and BSC Channels, EPFL

Fall 2014-2015

- *Supervisor*: Prof. Emre Telatar.
- *Subject*: Developed alternative proof for strong resolvability for the BEC and BSC channels: for a given channel and a probability distribution P, we find the minimum rate at which we need to encode the input so that the probability distribution of the output is arbitrarily close to P.

Security in Information Theory, EPFL

Fall 2013-2014

- *Supervisor*: Prof. Emre Telatar.
- *Subject*: Studied confidentiality in the wiretap and broadcast channels based on the works of *A.D. Wyner* and *Csiszár and Körner.*

Multi-access Channel Capacity Under Maximal Error Probability Criterion, EPFL

Spring 2012-2013

- *Supervisor*: Prof. Emre Telatar.
- *Subject*: Developed a revised proof for the multiple access channel capacity region formula as well as a detailed revision of Dueck's paper, *Maximal Error Capacity Regions Are Smaller Than Average Error Capacity Regions for Multiuser Channels.*

Planning LTE E-HNB Network Tool (PLENT), AUB, EPFL

2011-2012
Final Year Project

- *Supervisor*: Dr. Zaher Dawy.
- *Subject*: Designed and implemented a cellular planning tool (PLENT) that takes the potential impact of femtocells on the network into consideration and provides the operator with an optimal deployment configuration for its base stations. This product aims at helping mobile operators to efficiently handle the introduction of femtocells into the market.

WORK EXPERIENCE

Sequans Communications, Paris, France – Hardware Algo Intern

02.2014 - 08.2014

- Increased the connection speed achievable with Sequans' chips at low SNR while reducing their per unit cost, thus making them more competitive on the market, by:
  - Reducing the complexity of the legacy MMSE equalizer through new optimal designs.
  - Upgrading the legacy MMSE equalizer to support 4x4 MIMO.
  - Optimizing the 4x4 MMSE equalizer to have minimum complexity.
  - Implementing the new designs in Sequans' LTE physical layer C-simulator.

Mobilonia, Beirut, Lebanon – Software Intern

07.2013 - 09.2013

- Redesigned the instant messaging mobile application MobiChat, rendering it more robust to failures, more attractive to the user and easier to deploy on the market by:
  - Introducing push notifications.
  - Cutting any dependencies of the deployment system on any auxiliary server by deploying the SQL database on a company server.
  - Redesigning the user interface for Android.

AUB, Beirut, Lebanon – Research Assistant

07.2012 - 08.2012

- Enhanced PLENT (Planning LTE E-HNB Network Tool) the work on which started during the Final Year Project, under the supervision of Dr. Zaher Dawy by:
  - Optimizing the MATLAB code.
  - Developing a new GUI for an easier interaction with the user.

Ericsson Silicon Valley, San Jose, California, USA – Software Intern

06.2011 - 09.2011

- Slashed the time spent by the company on code optimization by 20% through:
  - Developing a full set of shell scripts to measure the code coverage of various unit-test scripts and functional smoke-tests.
  - Developing a full set of RTD (Routing Test Daemon; a locally developed Unit-Test) scripts to identify performance bottlenecks for the Routing Information Base (RIB) sub-system.

Ericsson Lebanon Communication, Beirut, Lebanon – Intern

08.2010 - 09.2010

- Worked on a case study analysis for Mobile Broadband for the African markets which helped clarify the company's policy to the employees.

TECHNICAL SKILLS

Wireless Communication

- Solid understanding of LTE, UMTS, GSM, WiMAX, theory and implementation.
- Working on planning tools for cellular networks such as TEMS.

Software Engineering

- Writing Shell scripts for Unix/Linux and configuring network devices such as routers, switches.
- Working with Git, CVS, SVN, CodeReview, IBM Extraview.
- Developing Android mobile applications.
- Database management and SQL.
- Solid understanding of Machine Learning principles and algorithms (NN, SVM, PCA...).

Programming Languages

- Python, C/C++, Java, Matlab, Labview, VHDL, PCspim (MIPS32 assembly language), Latex, Cadence, Sage, Microsoft Office.

LANGUAGES

Arabic	Native language
English	Fluent spoken and written
French	Fluent spoken and written
German	Intermediate level, B1 (European scale)

HONORS AND AWARDS

EPFL

- (2014) Receiver of EDIC Fellowship ($51000) for PhD studies.
- (2012) Receiver of the Excellence Scholarship ($32000) for excellent performance in undergraduate studies.

AUB

- (2008) Receiver of the Merit full scholarship from the American University of Beirut which is given to the top 10 applicants in Lebanon ($72000).

High School

- (2008) First prize for best philosophical dissertation ($700).
- (2008) First prize for best performance in Maths and Science ($700).
- (2007) Jacques Eddé prize for best performance in History/Geograpghy ($200).

PUBLICATIONS

- E. Najm and R. Nasser, "Age of information: The gamma awakening," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 2574–2578.
- E. Najm, R. Yates, and E. Soljanin, "Status updates through M/G/1/1 queues with HARQ," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 131–135.
- R. D. Yates, E. Najm, E. Soljanin, and J. Zhong, "Timely updates over an erasure channel," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 316–320.
- E. Najm and E. Telatar, "Status updates in a multi-stream M/G/1/1 preemptive queue," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, April 2018, pp. 124–129.
- E. Najm, R. Nasser, and E. Telatar, "Content Based Status Updates," in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 2266–2270.

**Ce document a été imprimé au Centre d'impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.**