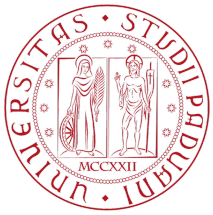

UNIVERSITY OF PADOVA

Department of Physics and Astronomy "Galileo Galilei"

Master's Degree in Physics of Data

Master's Thesis

Information freshness under non-terrestrial connectivity

Master's Student	**Lucrezia Rossi**
Supervisor	**Leonardo Badia**
Co-Supervisor	**Andrea Munari**

Academic Year 2024/2025

Abstract

Information *freshness* is becoming increasingly critical with the proliferation of communication networks and portable communication devices. Non-terrestrial networks exhibit unique challenges arising from line-of-sight constraints, dynamic traffic conditions, and the diverse architectural components. These characteristics make them a compelling scenario for analyzing optimal configurations and energy consumption.

In this work, we examine two different non-terrestrial communication models: a deterministic model with intermittent connectivity caused by periodic loss of line-of-sight and a multi-user multi-queue model in which traffic is split over multiple servers. For the intermittent connectivity model, we derive a closed-form expression for the metric known as *Age of Information* and demonstrate its linear dependence on the generation rate, service rate and duration of the active phase, as well as its inverse relationship with the duration of the inactive phase.

We also introduce the concept of *packet usefulness* for the multi-user multi-queue scenario and derive a closed-form expression for the shared queue. Our results highlight the importance of traffic splitting in reducing both the Age of Information and the overall resource consumption.

Contents

1 Introduction

This chapter provides the introduction and context for this study. Section 1.1 discusses the emergence of real-time communications and the growing importance of timely data delivery in modern applications. Section 1.2 presents the concept of Age of Information, which formalizes the notion of information freshness. Section 1.3 reviews the relevant literature on intermittently available channels and multi-user/multi-queue systems, highlighting existing approaches and identifying gaps that motivate this work. Finally, Section 1.4 summarizes the contributions of this thesis, introducing the new analytical models developed and their significance for non-terrestrial networks and real-time communication systems.

1.1 Real-time communications

Real-time communications are characterized by constraints in the period in which data availability and management became critical to make fast and efficient decisions. During history, the human necessity to communicate has always been limited by constraints on distance, time, and modes of transmission. Improving the ability to send data in a timely way is a central challenge in modern communication networks.

As observed by David S. Alberts and Daniel S. Papp[1]:

"Successfully responding to these challenges will require three things. First, we will need to recognize that something has changed. Second, we will need to understand the implications of this change. Third, we will need to develop timely and effective responses."

Having access to up-to-date information in real time is fundamental in many scenarios. For example, being rapidly informed of the arrival of a tsunami can save lives thanks to timely warning systems; updated data on the river levels can prevent floods or help manage evacuations; monitoring in real time the position and the velocity of an airplane or a ship helps to avoid incidents. In all these scenarios, having the data is not sufficient: their timeliness, i.e., their *freshness* decides how much effectively the system can react to changes and optimize its performance.

It is exactly from this necessity to understand and quantify the continuous updating of information that arose the need to create theories on stochastic and dynamic systems, to gain a deeper understanding of how data packets are transmitted within a server.

Queueing theory is widely used to model waiting systems and stochastic flows of packets (or clients) and, together with Shannon's information theory [2][3], it is used to evaluate the capacity of the channel and the delay. These theories give us the mathematical tools to describe the waiting times, congestions, and flow dynamics by analyzing channel latency and throughput.

However, the ever-increasing use of communication networks and connectivity via portable devices has made it necessary that real-time status updates are as timely as possible. This makes previous metrics no longer sufficient and opens the way to the more recent concept of *Age of Information* [4][5][6], which formalizes in a more quantitative way how *fresh* the information is in a system.

The study of Age of Information is not limited just to theory: it has become essential to understand the behavior of real-time systems, evaluate performance, and develop optimization strategies that can reduce delays and inefficiencies. The purpose of this thesis, starting from consolidated mathematical

principles, is to model and explore the value of the information update in dynamic and complex scenarios.

1.2 Age of Information

The concept of *Age of information* (AoI) was formalized around 2011. In [7] and [8], the authors move beyond the notion of status update age and explicitly adopt the term *system age of information* to describe the timeliness of information at the receiver. Traditional performance metrics such as delay, throughput, and sampling rate were no longer sufficient to quantify the freshness of information at the receiver, so researchers began to look for a new and effective metric.

In 2012 Kaul et al. published [4], which marks the birth of the AoI metric as we know it today. AoI measures how old the most recently received update is with respect to its generation time, capturing the staleness of information rather than the transmission delay alone.

The way in which this metric describes the process is very simple: every time a monitor receives an update at time t, this contains a time-stamp $u(t)$, capturing the time at which information was generated, i.e. the *status age*. The difference between t and $u(t)$ is the *age* of the quantity monitored.

$$\delta(t) = t - u(t)$$

AoI grows linearly, and at every update it will be reset. For this reason, the AoI graph has a particular shape and is called a sawtooth graph (see Figure 2.1).

Knowing the age allows us to analyze models and understand in which situations real-time systems perform better (the lower the age, the better). There are two main ways to do an AoI analysis:

- Time-average Age
- Peak Age

The time-average AoI is defined as the long-term average of the age process and characterizes its steady-state temporal behavior. The second metric focuses on the peak values [9] of AoI and was introduced to simplify the mathematical analysis of the first. In fact, the average peak age and average age are not so far from each other in model investigation [10] and the peak age may also be a powerful tool in applications with a threshold restriction on *age*.

In our analysis, we focus on the time-average age for a better understanding of our models.

1.3 State-of-the-Art

The rise of IoT (Internet of Things), cyber-physical systems, and autonomous networks, where timely information is essential, has increased the importance of freshness [11][12]. For ambient monitoring, surveillance, automation, and any scenario in which timely awareness of events is important, fresh information must be used to avoid errors.

In queueing systems, the AoI metric highlights how congestion and scheduling policies affect information freshness. For example, Inoue et al. [6] show its usefulness in analyzing single-server queues under different service disciplines. Similarly, the survey by Yates et al. [5] provides a comprehensive analysis of AoI in multi-user M/M/1 queues, illustrating how queue length and service policies influence the timeliness of information.

The application of AoI to communication channels is a more recent development. In real networks, transmission is not always continuously available; interruptions can occur due to random transmission errors, such as erasures [13] or collisions [14][15]. However, scenarios in which connectivity is systematically limited by the physical characteristics of the communication medium (e.g., satellite line-of-sight restrictions) have been less explored. Badia and Munari [16] also consider the freshness of the information from an intermittent link due to satellite flybys, but the process follows a scheduled

pattern and the duration of the missing line-of-sight of the satellite is modeled through a Markov chain.

Instead, a scenario in which cyclic intermittent connectivity [17] is considered with a periodic pattern has never been examined before, possibly due to the difficult tractability of deterministic time intervals.

Another important topic for non-terrestrial communication is the analysis of a multi–user multi–queue system. Related works address these challenges separately. Moltafet et al. [18] and Yates et al. [19] analyze the problem of multi-user systems. The former performs an analysis based on the AoI metric to evaluate the average AoI for only one of the multiple users, deriving a closed-form expression but already so complex even if they have one server. In the second case, they used an SHS approach to arrive at a similar solution.

Bhati and Baze [20] studied a two-queue system in which the user sends a message with probability α to one queue and $1 - \alpha$ to the other. The final formula of the average AoI involves high-degree multinomials, so they searched for an approximation. By using a Gamma distribution, they approximate the system to a single-server M/M/1 queue and evaluate the average AoI showing the benefits of having more servers.

In any case, each of these two problems is already highly complex when studied independently, and to the best of our knowledge, no published work has addressed their combination. In this study, our aim is to go further by providing a framework that opens new directions for the study of this joint problem. This is particularly relevant in satellite communications, where the growing number of users and the increasing availability of multiple satellites per user make such models especially important.

Both intermittent connectivity and multi-user/multi-queue problems pose key challenges in modern communication networks. In particular, non-terrestrial networks (NTNs) [21] exhibit predictable coverage cycles and must support multiple users sharing limited communication resources, motivating the study of these two problems.

For this reason, the topics are addressed separately, but both constitute essential extensions of the existing literature on AoI.

1.4 Contributions

The purpose of this work is to study new scenarios that are important for NTNs. Develop a new analytical representation to answer the challenges rising in real-time systems, understand the behavior of more complex transmissions, and find the better strategy to optimize the connectivity.

Satellites at lower altitudes are not always available because of their orbital motion: they rotate around the globe, and the transmitter can see them just during precise windows. The intermittent connectivity model has the precise purpose to understand how much transmission is delayed in these situations and to completely analyze what happens during on/off phases of different duration. We develop an analytical model with deterministic cyclic intermittent connectivity (Chapter 3), which has not been studied previously. This problem can also be related to IIoT (Industrial Internet of Things), industrial systems are affected by energy or multitasking constraints[22][23][24]. This does not allow the links to be always available, even if they cannot have prolonged outages for the operability of the system.

We extend the analysis to a multi-user multi-queue system (Chapter 4) to answer the growing necessity to send messages as fast as possible. Different users can use only a server to send their messages, but this can create congestion and delay. To optimize the transmission, a user can send the remaining of his data to another satellite and optimize the communication. The impact of this choice on the average AoI is a very interesting point of discussion in both NTNs and traditional networks.

We also provide simulation results to validate the theoretical models and discuss potential strategies for scheduling and updating policies to minimize AoI in practical scenarios.

2

Background and Preliminaries

This chapter presents the theoretical background used in this thesis. Section 2.1 introduces the average Age of Information. Section 2.2 reviews the fundamentals of queueing theory, focusing on M/M/1 model relevant for AoI analysis. Section 2.3 introduces the Birth-Death Processes important for understanding queue dynamics. Section 2.4 provides an overview of NTNs and their implications for intermittent connectivity and multi-user/multi-queue systems.

2.1 Average Age of Information

The AoI is a metric that quantifies the timeliness of status-update systems. It evaluates how much time is passed since the most recent update generated by a sender has been received by the monitor or receiver. Let denote as t_n the sequence of generation times of our packets of data and t'_n the corresponding delivery times. The instantaneous age process is defined as:

$$\delta(t) = t - u(t),$$

where $u(t)$ is the timestamp of the most recently delivered update at time t. Between two updates the AoI increases linearly and after a delivery it drops to the system time of the received packet.

Now we want to define, as a metric important in our study, the average AoI Δ [5]:

$$\Delta = \lim_{T\to\infty} \frac{1}{T} \int_0^T \delta(t)\, dt.$$

From the previous consideration, the AoI evolves following a typical sawtooth graph as shown in Figure 2.1. For the nth delivered update, we define the time between the generation of two consecutive packets, called the inter-arrival time, as $I_n = t_n - t_{n-1}$ and the time between the generation and delivery of a packet, called the system time, as $T_n = t'_n - t_n$. The key idea in the graphical method of age analysis is to decompose the area defined by the integral into a sum of trapezoidal areas. As we can see, these regions can be expressed as the difference between two triangular areas.

$$Q_n = \frac{1}{2}(I_n + T_n)^2 - \frac{1}{2}T_n^2 = \frac{1}{2}I_n^2 + I_n T_n.$$

If the system is *ergodic*, time averages converge to the statistical average over the set of all realizations. In particular, as $T \to \infty$, the integral of the age process can be approximated by the sum of the areas of individual sawteeth divided by the total observation time:

$$\frac{1}{T}\int_0^T \delta(t)\, dt \approx \frac{\sum_{n=1}^{N(T)} Q_n}{T},$$

where $N(T)$ is the number of updates delivered up to time T. Using the law of large numbers, in the long-term limit we can replace the sum with expected values:

$$\lim_{T\to\infty} \frac{\sum_{n=1}^{N(T)} Q_n}{T} = \lim_{N\to\infty} \frac{\sum_{n=1}^{N} Q_n}{\sum_{n=1}^{N} I_n} = \frac{\mathbb{E}[Q_n]}{\mathbb{E}[I_n]}.$$

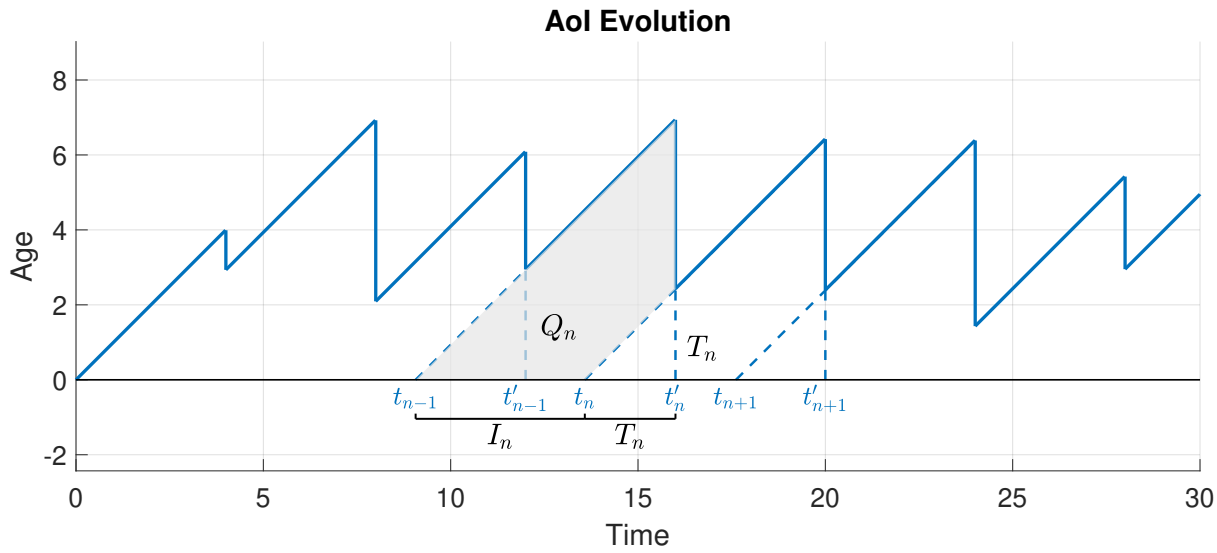

Figure 2.1: Sawtooth graph of the AoI as a function of time with focus on the nth inter-update interval.

Substituting the expression of Q_n in terms of I_n and T_n, we finally obtain:

$$\Delta = \frac{\mathbb{E}[Q_n]}{\mathbb{E}[I_n]} = \frac{\frac{1}{2}\mathbb{E}[I_n^2] + \mathbb{E}[I_n T_n]}{\mathbb{E}[I_n]}. \tag{2.1}$$

This equation can be applied to a broad class of service systems, such as FCFS (First-Come-First-Served) lossless systems and LCFS (Last-Come-First-Served) systems in which updates are preempted and discarded. It makes no specific assumptions regarding other traffic that might share the system with the update packets of interest. However, AoI analysis can be challenging, as evaluating the term $\mathbb{E}[I_n T_n]$ is difficult due to the negative correlation, which depends on the system dynamics.

However in scenario where the inter-arrival and the system times are independent, for example in bufferless or loss systems (no-queue) or in infinite-server systems, the expected value of $\mathbb{E}[I_n T_n] = \mathbb{E}[I_n]\mathbb{E}[T_n]$ and the average AoI became:

$$\Delta = \frac{1}{2}\frac{\mathbb{E}[I_n^2]}{\mathbb{E}[I_n]} + \mathbb{E}[T_n]. \tag{2.2}$$

This is used in the intermittent-connectivity model, as we want to analyze in greater detail how connectivity interruptions affect packet delivery. In contrast, in the multi-user/multi-queue system we have to use equation (2.1), since inter-arrival and service dynamics are influenced by the interaction between users and queues. However, since the term $\mathbb{E}[I_n T_n]$ cannot be computed analytically due to the complex interactions between users and queues, we restrict our study to simulation-based results. In general, the time-average AoI metrics provide a quantitative measure of information freshness and the associated formulas will be widely used in this work.

2.2 Queueing Theory Basics for AoI Analysis

In order to analytically derive the AoI, it is necessary to model the underlying process of packet generation, transmission, and service time. In this work, we will use queueing theory results to this aim, which are briefly introduced in this section.

Queueing theory was developed to model any type of traffic, whether composed of people or data packets, and to know how to optimize it. The average number of customers, the average amount of time a customer spends waiting in the queue, in which case congestion happens: these are all quantities that queueing system models can estimate.

In general, we describe the type of queue system with Kendall notation:

$$A/S/c/K/N/D$$

where A is the arrival distribution, S is the service time distribution, c is the number of servers. K represents the maximum number of customers that can be in the system and N is the population size, but they are usually omitted, considered infinite. D is the service discipline and, by default, is FCFS unless otherwise specified.

Examples of the more studied model are:

- M/M/1: Poisson arrivals, exponential service time, 1 server
- M/D/1: Poisson arrivals, deterministic service time, 1 server
- M/G/1: Poisson arrivals, general service time distribution, 1 server
- G/M/1: general distribution of arrivals, exponential service time, 1 server

Principally we see the use of Poisson processes and exponential variables, because in an already complex problem it is more easy to manipulate exponential quantities.

Moreover, queueing strategies also depend on how the service is managed. This can be:

- *Non-preemptive*: once a customer begins service, it cannot be interrupted until completion;
- *Preemptive*: the service of a customer can be interrupted by the arrival of another customer according to priority rules.

In this work, we focus on FCFS non-preemptive systems.

2.2.1 M/M/1 queue

The M/M/1 model represents the easiest and at the same time most fundamental among the models of queueing theory. It describes a system with a single server where the arrivals follow a Poisson process with inter-arrival times exponentially distributed with parameter λ and the service times are distributed exponentially with parameter μ.

The model denotes a process of birth and death, with state space $\{0, 1, 2, \ldots\}$ where each state corresponds to the number of clients in the system at a given time, including those in service.

The system is stable if the number of customers in the system does not grow to infinity over time, and in an M/M/1 queue this happens if and only if $\lambda < \mu$. In this case, the system at any part in time is in stationary conditions and follows a geometric distribution, so that its probability is:

$$P(\text{number of customers in the system} = n) = \pi_n = (1 - \rho)\rho^n, \tag{2.3}$$

where $\rho = \lambda/\mu$ is the utilization of the buffer. From this distribution, we can immediately evaluate that the average number of clients in the system is equal to $\rho/(1 - \rho)$ and its variance $\rho/(1 - \rho)^2$. These results, in closed-form, show the analytical power of the model and his capacity to immediately show the behavior of the system.

Another central aspect of this system is about the inter-arrival time and system time. The inter-arrival time is described by a Poisson process, so its probability density function follows an exponential distribution $\lambda e^{-\lambda t}$. From this distribution, the moments of the inter-arrival time can be directly obtained. In particular, the average inter-arrival time I is given by:

$$\mathbb{E}[I] = \frac{1}{\lambda}, \qquad \mathbb{E}[I^2] = \frac{2}{\lambda^2}. \tag{2.4}$$

In turn, the average system time is equal to $1/(\mu - \lambda)$. To be more precise, system time T is composed of two components: the time that a client spends in line, waiting time W, and the time needed to be processed by the server, service time S.

Under FCFS non-preemptive service, a newly arrived packet must wait for all customers ahead to complete service. If n customers are already in the system, the total service time before completion is the sum of $(n+1)$ independent exponential random variables with parameter μ. This sum follows an Erlang distribution, whose cumulative distribution function (CDF) is:

$$F(x; k+1, \mu) = 1 - \sum_{n=0}^{k} \frac{1}{n!} e^{-\mu x} (\mu x)^n. \tag{2.5}$$

Then, the CDF of the system time T can be written as a weighted sum of Erlang distributions:

$$F_T(x) = \sum_{n=0}^{\infty} \pi_n F(x; n+1, \mu) = \sum_{n=0}^{\infty} (1-\rho)\rho^n \left[1 - \sum_{k=0}^{n} \frac{(\mu x)^k}{k!} e^{-\mu x} \right]. \tag{2.6}$$

This gives exactly the system time distribution for an M/M/1 queue, which is equivalent to an exponential distribution with rate $\mu - \lambda$:

$$F_T(x) = 1 - e^{-(\mu-\lambda)x}, \quad x \geq 0.$$

The average AoI for an M/M/1 queue can also be exactly evaluated, resulting in a simple closed-form expression [5]:

$$\Delta_{\mathrm{M/M/1}} = \frac{1}{\mu} \left(1 + \frac{1}{\rho} + \frac{\rho^2}{1-\rho} \right). \tag{2.7}$$

This formula captures the freshness of information in the system, showing that AoI depends not only on the service rate but also on the system utilization ρ.

In summary, the M/M/1 model is an essential starting point in the study of queues and its simplicity allows us to obtain closed-form and easy to evaluate formulas. Its generality also make it a reference point to understand more complex phenomena.

2.3 Birth-and-Death Processes

A birth-death process is an important class of continuous-time Markov chains with countably infinite state space in which transitions are only allowed between neighboring states [25]. Let $X(t)$ denote the state of the process at time $t \geq 0$. The dynamics of the process are fully characterized by two sequences of non-negative parameters: the *birth rates* $\{\lambda_n\}_{n\geq 0}$ and the *death rates* $\{\mu_n\}_{n\geq 1}$, where:

- a *birth* corresponds to a transition $n \to n+1$,
- a *death* corresponds to a transition $n \to n-1$.

By construction, transitions to negative states are not allowed. For a small time interval $\Delta t > 0$, the transition probabilities satisfy:

$$\begin{aligned} P_{n,n+1}(\Delta t) &= \mathbb{P}\{X(t+\Delta t) = n+1 \mid X(t) = n\} = \lambda_n \Delta t + o(\Delta t), \quad n \geq 0, \\ P_{n,n-1}(\Delta t) &= \mathbb{P}\{X(t+\Delta t) = n-1 \mid X(t) = n\} = \mu_n \Delta t + o(\Delta t), \quad n \geq 1, \\ P_{n,n}(\Delta t) &= 1 - (\lambda_n + \mu_n)\Delta t + o(\Delta t), \quad n \geq 1. \end{aligned}$$

These relations explicitly express the idea that, if the interval is infinitesimal, at most one transition can occur and only with adjacent states.

The process can also be represented in terms of its *infinitesimal generator matrix* (or transition rate matrix) $Q = [q_{ij}]_{i,j\in\mathcal{S}}$, whose entries are defined as:

$$q_{n,n+1} = \lambda_n, \quad q_{n,n-1} = \mu_n \ (n \geq 1), \quad q_{n,n} = -(\lambda_n + \mu_n),$$

with all other entries equal to zero. Its tridiagonal structure reflects the fact that transitions occur only between neighboring states:

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \cdots \\ 0 & 0 & \mu_3 & -(\lambda_3+\mu_3) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The evolution of the probability distribution of the system state is governed by the Kolmogorov forward equations. Let $P_n(t) = \mathbb{P}\{X(t) = n\}$ denote the probability that the process is in state n at time t. Then, the system is described by the following differential equations:

$$\begin{aligned} \frac{d}{dt}P_0(t) &= \mu_1 P_1(t) - \lambda_0 P_0(t), \\ \frac{d}{dt}P_n(t) &= \lambda_{n-1}P_{n-1}(t) + \mu_{n+1}P_{n+1}(t) - (\lambda_n + \mu_n)P_n(t), \quad n \geq 1. \end{aligned}$$

Equivalently, we can write them in vector-matrix form and defining $\mathbf{P}(t) = [P_0(t), P_1(t), \dots]^T$, we have:

$$\frac{d}{dt}\mathbf{P}(t) = Q\mathbf{P}(t).$$

The formal solution of this expression is:

$$\mathbf{P}(t) = e^{Qt}\mathbf{P}(0),$$

where $\mathbf{P}(0)$ is the initial probability distribution and e^{Qt} denotes the exponential matrix. This expression allows us, in principle, to compute the probability of having any number of individuals (or packets) in the system at any time t.

A particularly important case of a birth and death process is the M/M/1 queue, where the birth and death rates are constant.

$$\lambda_n = \lambda, \quad \mu_n = \mu, \quad \forall n \geq 0 \ (\mu_0 = 0).$$

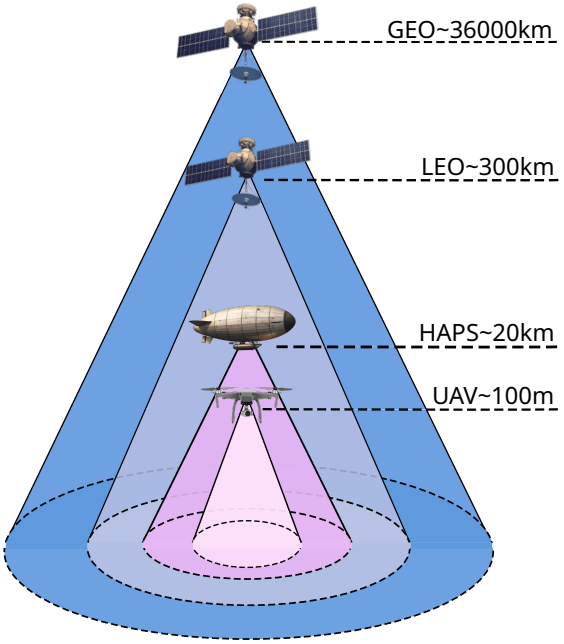

Figure 2.2: Non-terrestrial station, recreated based on [21].

This is very important for our analysis, as it provides a rigorous framework to model the dynamics of queues and to estimate the probability that a packet finds a queue empty or congested. In the multi-queue, multi-user scenario, the birth–death process is particularly relevant, since the system can be decomposed into M/M/1 queues.

2.4 Non-Terrestrial Networks

Non-terrestrial networks (NTNs) are radio communication systems operating far above the Earth's surface. They can include (see Figure 2.2) satellites in the lower terrestrial orbit (LEO), medium orbit (MEO), geostationary orbit (GEO), high altitude platforms (HAPs) and unmanned aerial vehicles (UAVs).

Recently, NTNs have been included in the standard 3rd Generation Partnership Project (3GPP), which allows them to operate as part of the 5G mobile network and makes them an integral part of the global mobile infrastructure [26].

This growing interest in non-terrestrial platforms is also reflected in recent survey literature, which highlights how satellite communications and related NTN technologies are becoming a key enabler for next-generation wireless systems. In particular, emerging applications such as 5G integration, remote sensing and global connectivity are driving renewed research efforts and large-scale deployments in space-based communication infrastructures, positioning NTNs as a crucial component of future wireless networks [27].

The principal components of NTNs are:

- a terrestrial terminal;
- an air or space station that can operate in a similar way as a terrestrial base station;
- a service link with the terrestrial terminal;
- a gateway that connects the non-terrestrial access network to the core network through a feeder link.

NTNs can adopt either a bent-pipe or a regenerative architecture. In a bent-pipe configuration, the satellite or aerial platform primarily forwards the signal to the gateway, which performs most of the processing. In a regenerative setup, some processing is carried out directly on the platform, potentially reducing delay and alleviating the gateway's workload. The gateway's role is therefore critical, as it influences latency, buffering and overall system performance [28].

UAVs or drones fly at a lower altitude and, thanks to their flexibility, are a good tool for broadband wireless connectivity. They can be deployed on-demand, allowing for rapid coverage of areas with high user density or inadequate infrastructure. HAPs operate in the stratosphere and have a wide geographic coverage of hundreds of kilometers. They are cheaper to operate with respect to terrestrial infrastructures, but may suffer from the need for refueling.

Compared to UAVs and HAPs, satellites provide more persistent connectivity and larger coverage areas, ensuring continuous service over extended durations. However, this comes at the cost of increased propagation delay and higher energy requirements. This means that system parameters need to be carefully configured to achieve optimal performance [27].

Satellites are also an integral part of NTNs and the primary focus of our study. In particular, GEO satellites (about 35 800 km of altitude) and LEO satellites (between 200 and 2 000 km). The former follow geodesic trajectories and cover very large geographic areas, making them always visible from terrestrial terminals. However, the long distance from Earth causes signal propagation delay and attenuation. LEO satellites, instead, guarantee better signal strength, but they are not stationary relative to the surface. For this reason, they are not always visible and must operate in constellation to maintain service continuity.

Compared to terrestrial networks, NTNs have unique challenges, such as variable latency, incomplete coverage, weaker signal, and strong Doppler shifts (rapid movements create a change in the frequency of the signal). However, with respect to terrestrial networks, NTNs offer wider geographic coverage, reduced congestion, and more predictable propagation conditions. These characteristics make NTNs both a challenging and valuable environment for examining the timeliness of information. The analysis of AoI can respond to this challenge and find in which context updates can be as timely as possible, and design strategies to optimize the freshness of information.

2.4.1 Traditional QoS Metrics vs Age of Information

Network performance is commonly evaluated using classical Quality of Service (QoS) metrics [29], such as:

- *Throughput*: the rate at which data is successfully delivered over the network.
- *Latency / Delay*: the time it takes for a packet to travel from the sender to the receiver.
- *Packet Loss*: the fraction of packets that do not reach the receiver.
- *Jitter*: the variation in packet arrival times.

Although throughput provides a measure of the network's capacity, it does not capture how timely the received information is. Similarly, latency and jitter quantify transmission delays and variability, but offer an instantaneous view that does not capture freshness. Packet loss measures reliability, but does not indicate whether successfully delivered packets are still relevant or fresh for the application.

Also, these metrics are suitable for terrestrial networks with relatively stable links. NTNs have characteristics that limit their effectiveness, as they often experience intermittent connectivity, frequent handovers, and time-varying link quality due to satellite mobility or orbital dynamics.

Consequently, packets may be delivered with significant delay or after temporary disconnections, making classical QoS metrics insufficient to evaluate the actual value of the received information. Thus, AoI provides a more application-oriented metric, measuring the time elapsed since the generation of the most recently received update and offering a direct assessment of information freshness.

3

Cyclic Intermittent Communication: Modeling and Information Freshness

This chapter introduces a cyclic intermittent connectivity model to describe communication systems with periodic service availability. Section 3.1 analyzes the packet transmission process under intermittent connectivity with asymmetric active/inactive phases, with a brief mention of the symmetric case. Section 3.2 extends the analysis to the scenario in which the system starts in the inactive phase. Finally, Section 3.4 evaluates the AoI.

3.1 Intermittent Transmission with Asymmetric Phases

The purpose of this analysis is to investigate the behavior of an intermittent transmission process in which the transmitter alternates between the active and inactive phases. This type of model is particularly relevant when describing satellite communication systems, where transmission opportunities depend on periodic visibility windows between the satellite and the terrestrial station (see Figure 3.1). The duration of these phases are, in general, different. In this part of the analysis, we concentrate on a process that starts in an active phase. When our packet is generated, the server cannot always send it to the receiver, because transmission is only possible during active windows. The process (see Figure 3.2) is characterized by exponential times and the service happens with rate μ only during active phases. The most important quantities for this analysis are:

- the duration of the active phase T_1;
- the duration of the inactive phase T_2;
- the delivery time F of each packet;
- the service rate μ.

We denote by $T = T_1 + T_2$ the total duration of one complete cycle, consisting of one active phase followed by one inactive phase.

This modeling framework is aligned with recent research showing the limitations of traditional metrics in satellite networks and the relevance of AoI as freshness metric. For example, in [30] a multi-hop LEO constellation subject to erasures and queueing delays is analyzed and are derived closed-form bounds for average and peak AoI. The study reveals non-trivial trade-offs between update rate, packet loss and network load. This strongly impact the freshness of data at the receiver despite potentially high throughput.

To evaluate the AoI, we use formula (2.2). The two quantities that we need to compute it are the expected value of the inter-arrival time $\mathbb{E}[I]$ and the expected value of the system time $\mathbb{E}[F]$. In this section, we estimate the second one.

The probability that the transmission is not completed by time f depends on which phase we are in. If f belongs to the jth active phase, the probability depends on two independent events: the transmission failure in the previous active phase and the failure until f during the current active phase. In contrast, if we are in the inactive phase, the probability p remains equal to one, since no

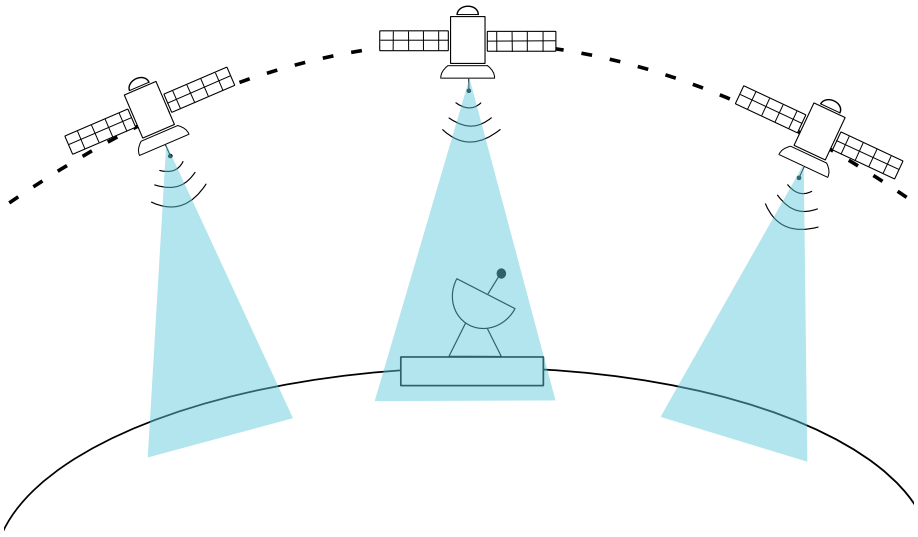

Figure 3.1: Illustration of intermittent satellite-to-receiver transmission. The receiver is only within the satellite's visibility window for limited periods, resulting in cyclic connectivity and periods of inactivity.

transmission can occur when the server is down (the packet will not be delivered).

$$p = \begin{cases} e^{-j\mu T_1} \cdot e^{-\mu(f-jT)}, & \text{if } f \in [jT, jT+T_1] \text{ (active phase)} \\ 1, & \text{if } f \in [jT+T_1, (j+1)T] \text{ (inactive phase)} \end{cases}$$

This expression represents exactly the complementary probability of the cumulative distribution function and, by deriving it, we can evaluate the probability distribution function (pdf). The pdf (see Figure 3.3) of the finishing time F can be written as:

$$p_F(f) = \begin{cases} \mu e^{-j\mu T_1} e^{-\mu(f-jT)}, & f \in [jT, jT+T_1] \\ 0, & f \in [jT+T_1, (j+1)T] \end{cases}$$

To compute the expected value $\mathbb{E}[F]$, we must take into account that we are considering deterministic time intervals. Therefore, to obtain the expected value in the jth interval, we integrate the pdf multiplied by the time f. The final value is given by summing over all possible intervals:

$$\mathbb{E}[F] = \sum_{j=0}^{\infty} \int_{jT}^{jT+T_1} f \cdot p_F(f)\, df = \sum_{j=0}^{\infty} I_j.$$

The I_j component is:

$$I_j = \int_{jT}^{jT+T_1} f \cdot \mu e^{-\mu(f-jT)} e^{-j\mu T_1}\, df = \left(\frac{1-(1+\mu T_1)e^{-\mu T_1}}{\mu} + jT(1-e^{-\mu T_1}) \right) e^{-\mu j T_1}.$$

Thus, we are left only with the infinite summation, where each term represents the contribution of the jth active phase to the overall expectation.

$$\mathbb{E}[F] = \sum_{j=0}^{\infty} e^{-j\mu T_1} \left[\frac{1-(1+\mu T_1)e^{-\mu T_1}}{\mu} + jT(1-e^{-\mu T_1}) \right]. \tag{3.1}$$

Although this expression is exact, it is not immediate to interpret or evaluate numerically. So, to obtain a more useful result, we can solve it by splitting our equation and using series summation identities. After defining auxiliary variables:

$$X = \frac{1-(1+\mu T_1)e^{-\mu T_1}}{\mu}, \quad Y = T(1-e^{-\mu T_1}),$$

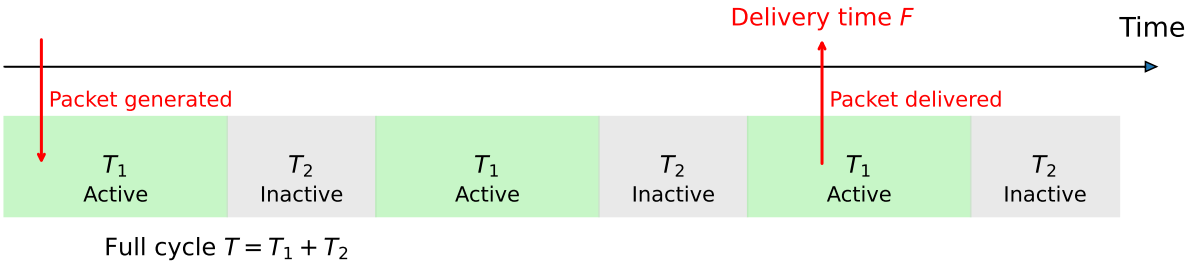

Figure 3.2: Intermittent satellite-to-receiver transmission. The transmitter alternates between active T_1 and inactive T_2 phases, and packets can only be delivered during active phases.

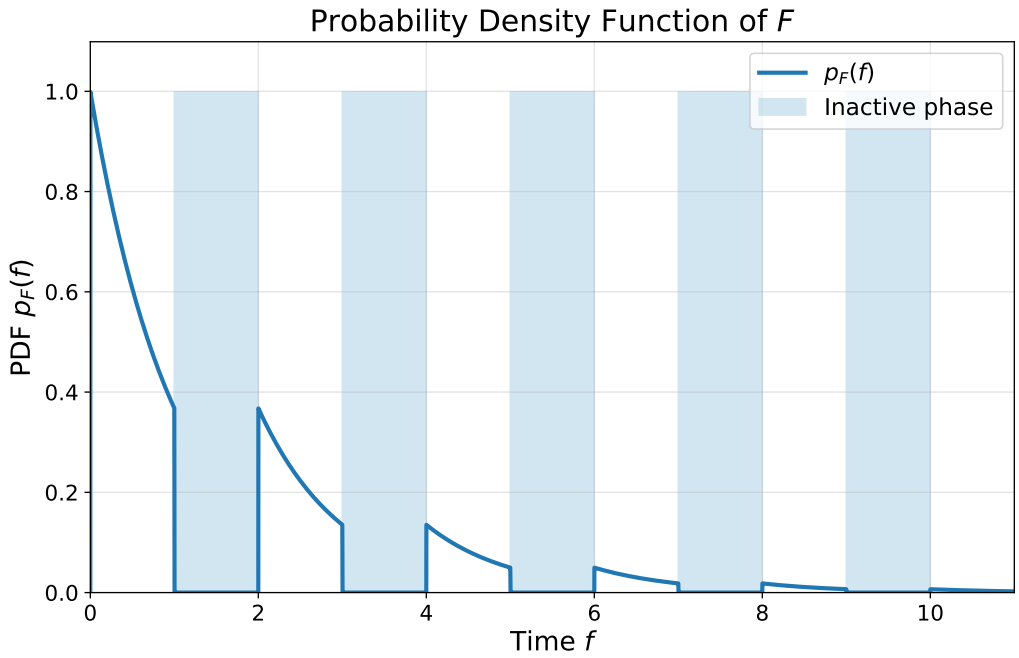

Figure 3.3: Probability density function $p_F(f)$ of the packet delivery time F in a scenario with intermittent transmissions. During inactive periods, no packets are sent to the receiver.

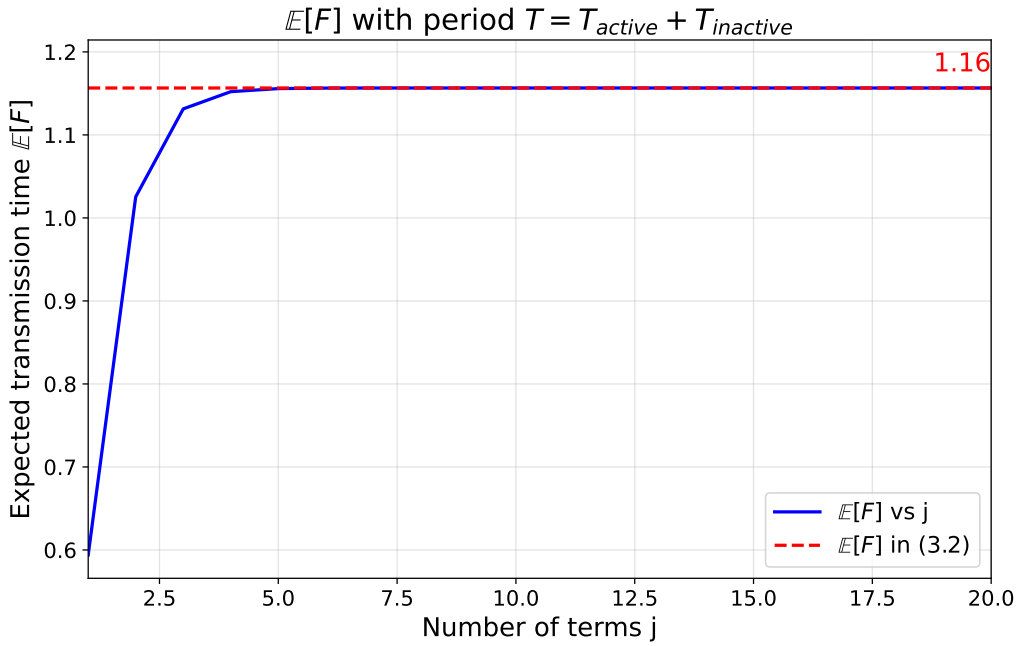

Figure 3.4: Expected transmission time for parameters $\mu = 1$, $T_1 = 2$, and $T_2 = 1$. The solid blue line shows the series-based computation, while the red dashed line represents the analytical result.

we get:

$$\begin{aligned}\mathbb{E}[F] &= X \sum_{j=0}^{\infty} e^{-\mu j T_1} + Y \sum_{j=0}^{\infty} j e^{-\mu j T_1} \\ &= \frac{X}{1 - e^{-\mu T_1}} + \frac{Y \cdot e^{-\mu T_1}}{(1 - e^{-\mu T_1})^2}.\end{aligned}$$

This result is obtained by using two standard series:

- Geometric series: $\sum_{j=0}^{\infty} e^{-j\mu T_1} = \frac{1}{1-e^{-\mu T_1}}$,
- Weighted geometric series: $\sum_{j=0}^{\infty} j e^{-j\mu T_1} = \frac{e^{-\mu T_1}}{(1-e^{-\mu T_1})^2}$.

After some simplifications, we obtain the final closed-form expression for the expected delivery time $\mathbb{E}[F]$, which explicitly depends on the service rate μ and the duration of the active and inactive phases T_1 and T_2.

$$\mathbb{E}[F] = \frac{1 - (1 - \mu T_2) e^{-\mu T_1}}{\mu (1 - e^{-\mu T_1})}. \tag{3.2}$$

This closed-form result gives us important insight for the system analysis. The overall performance is driven by both the service dynamics and the structure of the satellite visibility process.
In Figure 3.4, we compare (3.1) and (3.2). This validates the correctness of the derived expression, since the two results coincide perfectly. Moreover, from the blue line we can identify the average number of active phases required to successfully deliver my packet to the final destination, giving us additional information about the transmission dynamics. Finally, for a complete analysis of the expected value, we evaluate how it changes with the durations of the active phase T_1 and the inactive phase T_2. As shown in Figures 3.5 and 3.6, the period of the inactive phase increases the average time delivery, while the active duration T_1 reduces it. This already gives us important insights on the delay of the transmission and the timeliness of the received information.

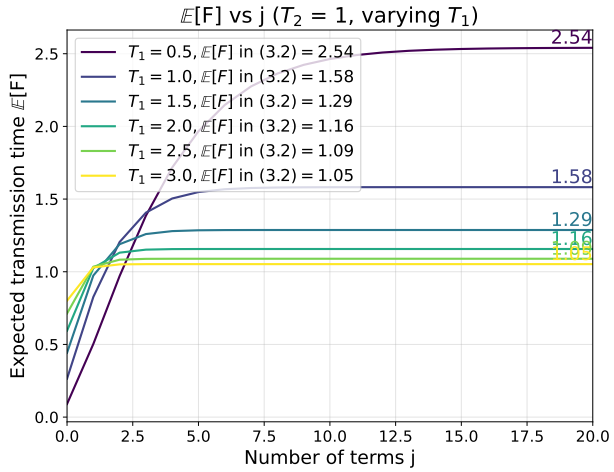

Figure 3.5: Expected transmission time for $\mu = 1$, $T_2 = 1$, and varying values of T_1.

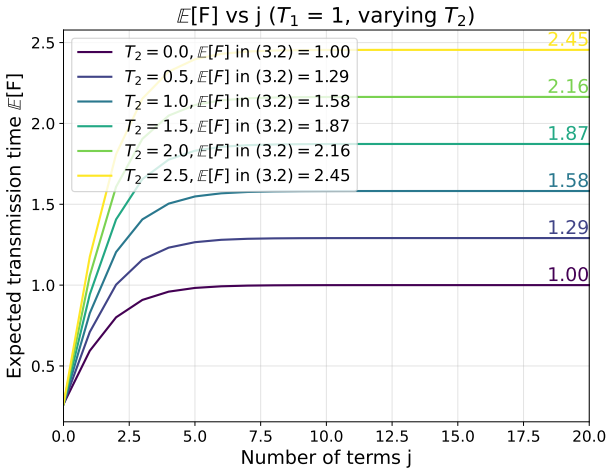

Figure 3.6: Expected transmission time for $\mu = 1$, T_1 = 1, and varying values T_2.

3.1.1 Symmetric Connectivity Phases

The symmetric transmission scenario, where $T_1 = T_2 = T$, is a special case of the previous. It yields several simplifications and exhibits more regular behavior. The expected transmission time reduces to the closed-form expression:

$$\mathbb{E}[F] = \frac{1 - (1 - \mu T)e^{-\mu T}}{\mu(1 - e^{-\mu T})}. \tag{3.3}$$

To fully investigate the impact of intermittent connectivity on communication dynamics, we analyze how the expected transmission time varies with the transmission rate μ, as shown in Figure 3.7. The results show an inverse relationship between the service rate and the average delivery time: increasing μ leads to a faster delivery process and thus reduces the time needed to complete the entire procedure. This behavior is consistent with intuition, since a higher service rate increases the probability of completion during the active phase and reduces the number of cycles that the packets need to go through.

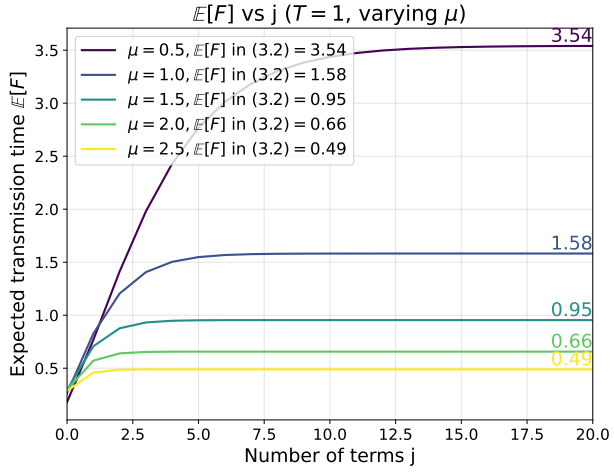

Figure 3.7: Expected transmission time $\mathbb{E}[F]$ as a function of the transmission rate μ for a fixed T. Increasing μ reduces the expected transmission time.

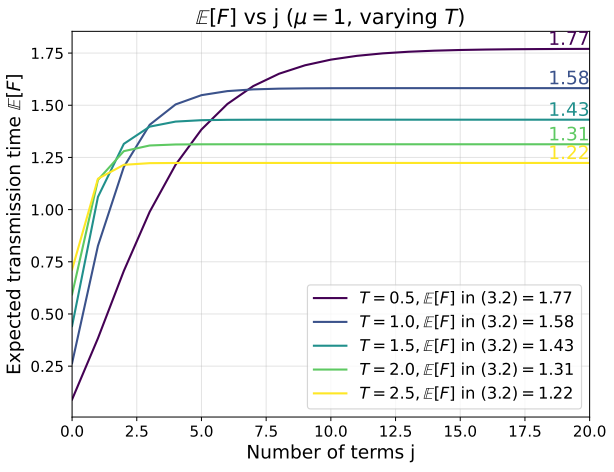

Figure 3.8: Expected transmission time $\mathbb{E}[F]$ as a function of the interval duration T for a fixed μ. Increasing T reduces the expected transmission time.

Moreover, Figure 3.8 reveals the impact of phase duration T on system performance. As T increases, both active and inactive intervals become longer, leading to longer waiting periods and an increase in overall system time. The symmetric configuration allows us to observe how the periodicity of the server availability alone, even in the absence of asymmetry, plays a fundamental role in determining the system delay.

3.2 Intermittent Transmission Starting in the Inactive Phase

To complete the analysis of the intermittent transmission model, we now investigate the case in which the system starts in an inactive phase. In this configuration, the system remains inactive for a duration T_2 during all odd intervals, which are then followed by an active phase of duration T_1 in the even intervals. This scenario is also relevant for practical applications since the satellite may initially be outside the visibility window and therefore unable to serve incoming packets.

As before, the transmission is only allowed during active intervals and the pdf of F becomes:

$$p_F(f) = \begin{cases} 0, & f \in [jT, jT+T_2) \\ \mu e^{-j\mu T_1 - \mu(f-(jT+T_2))}, & f \in [jT+T_2, (j+1)T) \end{cases}$$

Repeating the same analytical steps of the previous section where the system starts in the active phase, we derive an estimation of the expected delivery time by integrating the pdf over all possible active intervals. This results in the following infinite series representation:

$$\mathbb{E}[F] = \sum_{j=0}^{\infty} e^{-j\mu T_1} \left[\frac{1 + \mu T_2 - (1+\mu T) e^{-\mu T_1}}{\mu} + jT(1 - e^{-\mu T_1}) \right].$$

Once again, this expression can be simplified and, after appropriate manipulation, we arrive at the closed-form expression:

$$\mathbb{E}[F] = \frac{1 + \mu T_2 - e^{-\mu T_1}}{\mu(1 - e^{-\mu T_1})}.$$

Observing this formula more closely, we see that compared to the scenario where the system starts in the active state, the presence of an initial inactive interval introduces an additional delay equal to T_2. Specifically:

$$\mathbb{E}[F]_{\text{inactive}} = \mathbb{E}[F]_{\text{active}} + T_2.$$

This shows that the system behavior in the two scenarios is identical once the initial inactive phase is over.

Moreover, it is interesting to analyze some limiting and special cases of the derived expression. When $T_2 \to 0$, the expected delay reduces to that of a memoryless exponential distribution $\mathbb{E}[F] = \frac{1}{\mu}$. This confirms the consistency of the proposed model with standard queueing theory results. In the symmetric case where $T_1 = T_2 = T$, the expression is further simplified and we obtain:

$$\mathbb{E}[F] = \frac{1 + \mu T - e^{-\mu T}}{\mu(1 - e^{-\mu T})}.$$

Finally, Figures 3.9 and 3.10 show how the expected finish time evolves with respect to the active phase duration T_1 and the inactive phase duration T_2, respectively. In particular, Figure 3.9 displays that increasing the duration of the active phase reduces the expected delivery time, since it increases the probability of finishing the service in a single cycle. Instead, Figure 3.10 displays the negative impact of longer inactive periods, which directly increase the waiting time before the service can even start. These results show how both the structure of the intermittent connectivity and the ordering of the phases play a crucial role in determining the performance of the system. They provide useful insights for the design and optimization of communication systems that work under periodic service constraints.

3.3 AoI Evaluation

To evaluate the AoI, we consider a stochastic model in which the generation of each new packet is triggered only after the successful delivery of the previous one. In particular, as illustrated in Figure 3.11, the inter-generation time I_n follows an exponential distribution and its timer starts immediately after the delivery instant of the preceding packet. As a result, the evolution of AoI can be described as a sequence of independent sawtooth-shaped cycles, each corresponding to a triangular area Q_n. Each Q_n quantifies the age accumulated during the nth update cycle and resets to zero upon the reception of the new update at the receiver.

More precisely, updates are generated according to a delivery-triggered Poisson process with rate λ. This means that, although the inter-generation times are exponentially distributed, the generation clock is reset at every delivery event, and not continuously running as in a standard Poisson arrival process.

Moreover, we are not considering a queueing system in this model, since there is at most one packet in the system at any given time. Consequently, the service (or transmission) time is statistically independent of the update generation process and does not influence the subsequent inter-generation time.

The server represents a satellite with alternating active and inactive phases, corresponding to visibility and non-visibility intervals, respectively. The system time of this process has already been evaluated in the previous sections of this chapter. From the calculation done before, we know that if we start in the inactive phase, all is shifted by a factor T_2, so we limited our analysis only in case the packet arrives in the active phase, as the results will be redundant.

In this scenario, the average AoI can be estimated using (2.2), since the inter-arrival time I and the system time F are independent random variables. So using (2.4) we obtain the following closed-form expression:

$$\Delta = \frac{1}{\lambda} + \mathbb{E}[F],$$

where $\mathbb{E}[F]$ is the mean system time under intermittent connectivity, as in (3.2).

To evaluate the truthfulness of this model, we implemented a Monte Carlo simulation that explicitly reproduces the stochastic generation and service processes described above. We compare the theoretical formula of the AoI with the empirical results obtained over a large number of simulations. This comparison allows us to verify the validity of our expression and the robustness of the model under different operating conditions.

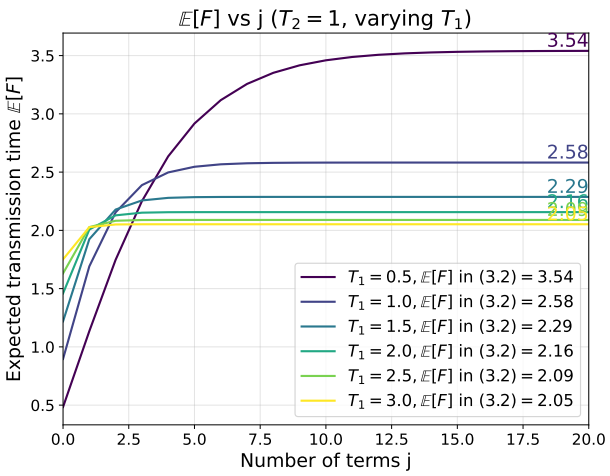

Figure 3.9: Expected transmission time $\mathbb{E}[F]$ as a function of T_1, for fixed $\mu = 1$ and $T_2 = 1$.

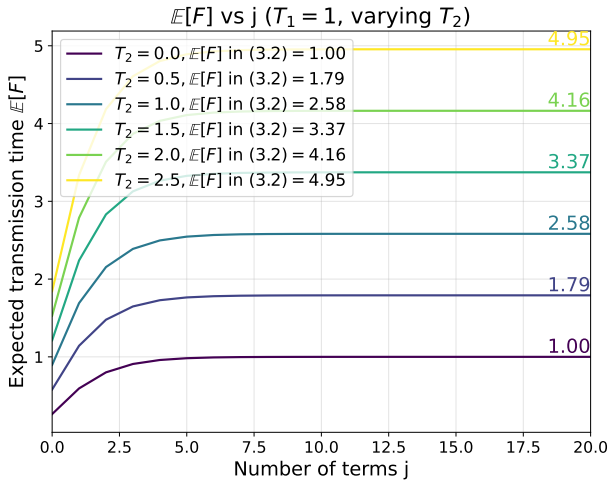

Figure 3.10: Expected transmission time $\mathbb{E}[F]$ as a function of T_2, for fixed $\mu = 1$ and $T_1 = 1$.

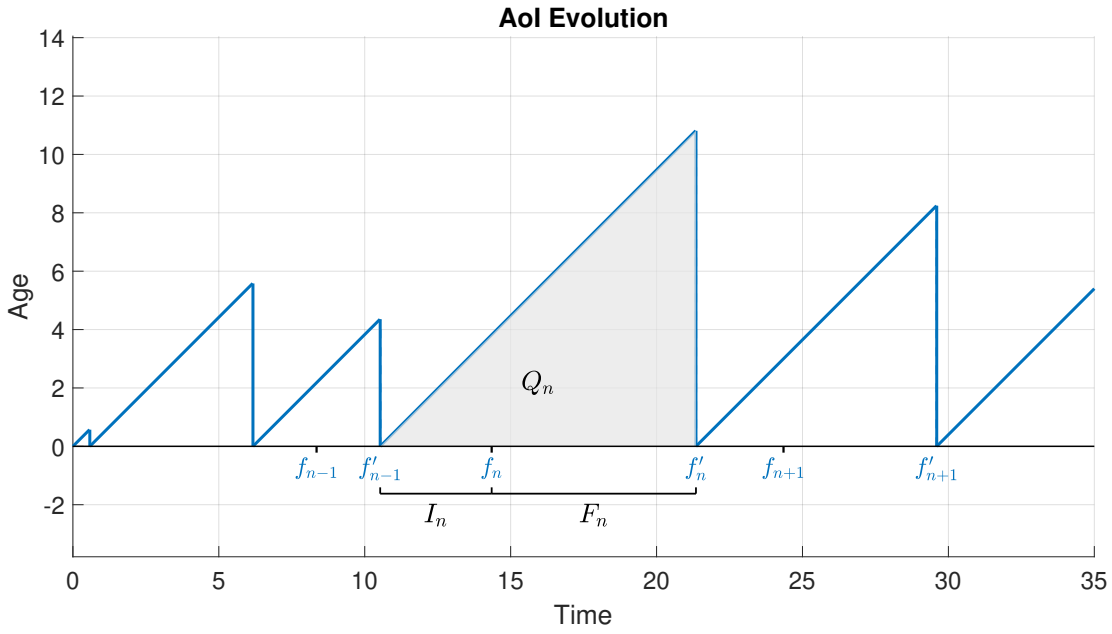

Figure 3.11: Evolution of the AoI over time. For the nth update, the inter-generation time I_n starts immediately after the reception of the previous packet, while the delivery interval F_n spans from the generation instant f_n to the reception instant f'_n. The triangular shaded areas Q_n represent the AoI contribution of each update cycle.

In particular, we decide to do separate study for each parameter of interest. These are the active phase T_1, the total period T, the rate of service μ and the rate of generation λ.

As shown in Figure 3.12, the AoI decreases significantly as the service rate μ increases, considering λ=0.8, T_1=2 and T_2=0. This behavior is intuitive, since faster servers are able to process packets more quickly, thus reducing delay. However, the reduction is non linear and already for $\mu > 3$ the improvements are marginal. This suggests that while the optimal AoI is achieved at higher service rates, a lower μ can be chosen to achieve a better trade-off between time and resource utilization.

A similar trend can be observed in Figure 3.13, where the average AoI decreases as the packet generation rate λ increases, with μ=1, T_1=2 and T_2=0. This implies that more frequent updates obtain fresher information. However, as with the service rate μ, the improvement decreases beyond a certain point, showing marginal gains for higher λ. This emphasizes the need to carefully select also an optimal generation rate to balance timeliness and system efficiency, rather than simply maximizing λ.

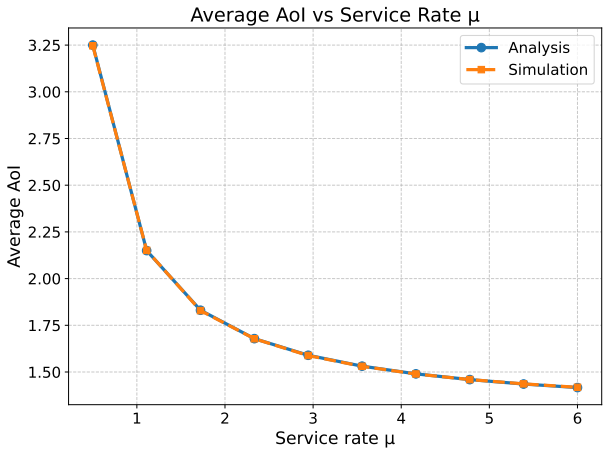

Figure 3.12: Average AoI as a function of the service rate μ, with λ=0.8, T_1=2 and T_2=0.

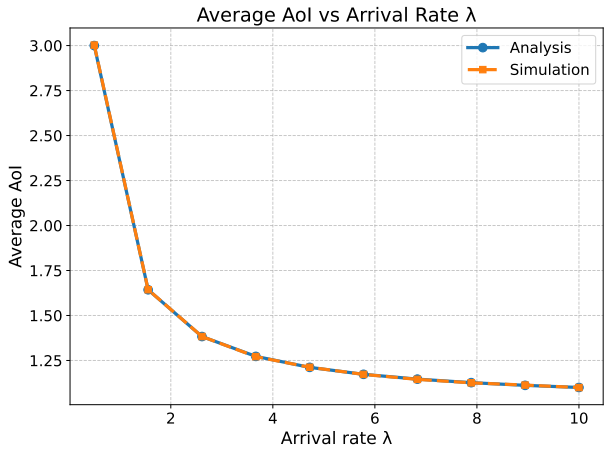

Figure 3.13: Average AoI as a function of the generation rate λ, with μ=1, T_1=2 and T_2=0.

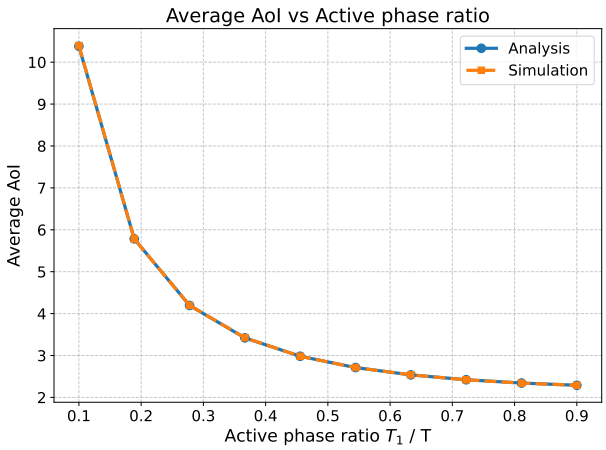

Figure 3.14: Average AoI as a function of the phase ratio T_1/T, with λ=0.8, μ=1 and T=2.

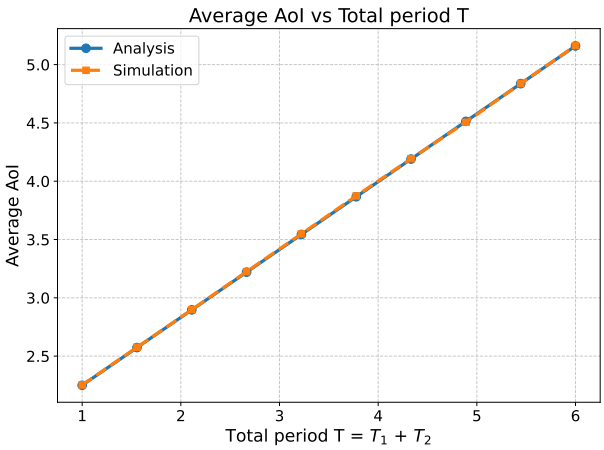

Figure 3.15: Average AoI as a function of the total period T, with λ=0.8, μ=1 and T_1=1

Figure 3.14 displays how AoI changes as a function of the ratio T_1/T, the fraction of time in which the system is in the active phase. We set λ=0.8, μ=1 and T=2. As expected, increasing this ratio improves performance because the system is available more frequently. From the figure, we can also see that the most significant improvement occurs for $T_1/T < 0.6$, after which the decrease in AoI becomes considerably less pronounced. This means that my transmission can be effective even if we have to transit in phases of unavailability due to physical or energetic constraints.

Finally, in Figure 3.15 we investigate the impact of the total period T while keeping T_1 fixed, with λ=0.8, μ=1 and T_1=1. In this case, increasing T means increasing the inactive phase T_2. The result show a linear correlation with the AoI, showing that longer inactive intervals lead to outdated information held at the receiver for longer time spans. These results provide important insight into real-world scenarios with intermittent connectivity. Connectivity patterns are often dictated by physical, technological, and energy constraints. The proposed analysis allows us to identify suitable operating points that balance AoI minimization with resource consumption, enabling the design of more efficient communication strategies.

4

Multi-User Multi-Queue System

This chapter presents a multi-user multi-queue system with traffic splitting. Section 4.1 introduces the dual-queue model and discusses its relevance for satellite-based communication systems. Section 4.2 develops the analytical derivation of the packet usefulness. Section 4.3 investigates the AoI performance through Monte Carlo simulation results and an age-energy trade-off.

4.1 System Architecture

The model of our interest is a system with M/M/1 queues FCFS without preemption (see Figure 4.1). The M/M/1 model is widely used to represent realistic scenarios such as distributed edge servers, uplink and downlink channels in LEO satellite networks [31], and shared servers in IoT and wireless sensor networks [32], where stochastic packet arrivals and service processes naturally arise. For this reason, it is an appropriate and well-established choice for modeling the system considered in this work.

Each user can rely on two queues: one just for himself and one shared with all other users. In general, we have N users which generate packets, each with rate λ. They send each newly generated packet either to their own queue, with probability α or to the shared queue, with probability $1 - \alpha$. The interesting point of this architecture is that a user can decide, to optimize his transmission, how much of his data to send to the shared queue.

The use of a routing probability α in our model is closely related to well-known mechanisms of probabilistic routing and randomized server selection in distributed systems. Similar approaches are widely used in the literature to achieve scalable and distributed load balancing, where each user makes an independent decision without requiring global knowledge of the system. In particular, random or probabilistic server selection has been shown to be effective in reducing congestion and improving overall system performance under strict latency constraints [33]. Moreover, classical results in randomized load balancing, such as the *power of two choices* paradigm and shortest-queue selection policies, have demonstrated that even simple randomized decisions can lead to significant performance gains in large-scale queueing systems [34]. Therefore, the introduction of the probability α is not an ad-hoc assumption, but rather a design choice grounded in established and well-studied principles in queueing theory and distributed network optimization.

The shared queue has a higher service rate μ_2 but its traffic may be high due to the presence of other clients. This setting closely resembles an edge–cloud (or LEO–GEO) continuum, where nearby resources such as edge servers or LEO satellites provide low-latency but limited processing capabilities, while remote and more powerful infrastructures such as cloud data centers or GEO satellites offer higher service rates at the cost of increased contention and queueing delays [35] [36]. Instead, local queues are dedicated to a single source, but offer a lower service rate μ_1. Which option is better for the user? He can decide to send all his data to only one of the two queues, but this decision can lead to a very high delay in transmission.

Given the symmetry of the problem, we focus on the transmission and the AoI with just one reference user. For simplicity, we reduce the model to just two queues: the private queue for user 1 and the shared queue that collects the aggregate traffic from all the other $N - 1$ users (see Figure 4.2). Since the other users are grouped together, the generation rate to the shared queue becomes $N(1 - \alpha)\lambda$. This simplification allows us to study the system in depth without losing generality.

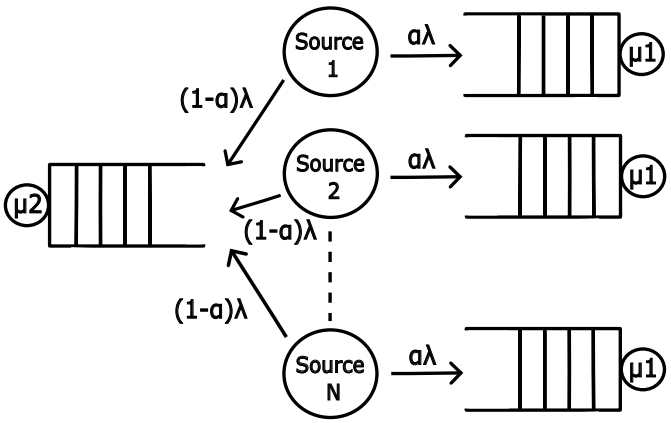

Figure 4.1: System model with N users, each having a private and a shared M/M/1 queue. Packets are directed to the private queue with probability α and to the shared queue with probability $1-\alpha$. All queues follow FCFS service without preemption.

This scenario may be relevant for NTN applications. In general, a user sends data to satellites with higher service rate μ, such as higher orbit satellites such as GEO, but they can be very busy due to heavy traffic from other clients. So, it can be advantageous to split the traffic and send data to a less capable but less congested server, such as a LEO satellite.

The optimal decision on how to split the traffic is not trivial and deserves an in-depth study. The overall behavior departs from the classical M/M/1 queuing system. In some cases, a packet delivered is no longer the *freshest*. To better understand, we define a packet as *useful* if it carries new information, and this implies that when it is delivered, it is not older than any previously delivered packets, i.e., it will reset the AoI for the user. The policy inside each queue is FCFS, so packets cannot overtake each other in the same queue. However, a packet in the opposite queue can, making one or more packets useless. Knowing whether the data provided to the collector are useful or not is a very important step in understanding under which conditions more packets are discarded. Assessing packet usefulness thus provides insight into the effects of queue interactions and traffic management, which directly impact AoI and energy efficiency. Transmitting packets that do not contribute to improving the system state leads to unnecessary consumption of energy and bandwidth, which is particularly critical in resource-constrained scenarios such as IoT devices, battery-powered sensors and satellite communications. Recent studies have highlighted this trade-off between information freshness and energy expenditure, showing that intelligent packet management can significantly reduce energy use while maintaining low AoI [37] [38].

4.2 Packet Usefulness

The purpose of this section is to analyze the *packet usefulness* for a single user. This metric allows us to distinguish between packets that effectively contribute to the information update process and those that become obsolete due to queue interactions.

To start our analysis, we define the main random variables associated with the traffic of User 1:

- T_1: system time of a packet entering queue 1;
- T_2: system time of a packet entering queue 2;
- I_1: inter-arrival time for packets entering queue 1;
- I_2: inter-arrival time for packets entering queue 2.

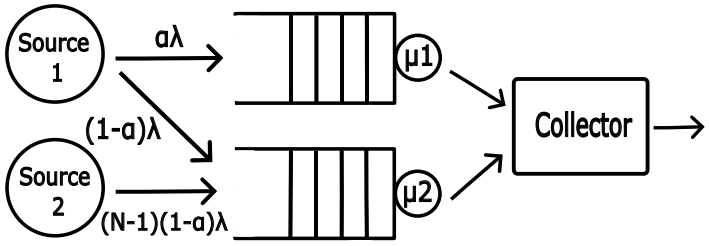

Figure 4.2: Reduced system model, showing a private M/M/1 queue and a shared queue that aggregates traffic from the other $N-1$ users. The total arrival rate to the shared queue is $N(1-\alpha)\lambda$.

Based on the known properties of M/M/1 queues (as discussed in Section 2.2.1), the probability density functions of these rvs are:

$$\begin{aligned} f_{T_1}(t) &= (\mu_1 - \alpha\lambda)\, e^{-(\mu_1 - \alpha\lambda)t}, \\ f_{T_2}(t) &= (\mu_2 - N(1-\alpha)\lambda)\, e^{-(\mu_2 - N(1-\alpha)\lambda)t}, \\ f_{I_1}(t) &= (\alpha\lambda)\, e^{-(\alpha\lambda)t}, \\ f_{I_2}(t) &= ((1-\alpha)\lambda)\, e^{-((1-\alpha)\lambda)t}. \end{aligned}$$

The final goal is to find the probability that a packet is useful. This occurs if and only if it is delivered before it is made obsolete by another, more recent, packet on the opposite queue. More precisely, a packet sent to queue 1 is useful if its system time T_1 is smaller than the time required for a newer packet from queue 2 to be generated and delivered. The same consideration is true for a useful packet in queue 2. So, the probability of being useful is defined as follows:

$$\begin{aligned} P_{1\,useful} &= P(T_1 < I_2 + T_2), \\ P_{2\,useful} &= P(T_2 < I_1 + T_1). \end{aligned} \tag{4.1}$$

To derive these quantities, we note that, although the two queues are independent, within the same queue the inter-arrival time and the system time of packets are not independent. In particular, the system time of the opposite queue change in consideration of packets already present in the queue upon its arrival, influenced by the departures occurring during the inter-arrival interval and arrivals only from other users. An illustrative situation is reported in Figure 4.3. At the arrival of a packet in queue 1, the shared queue contains n packets. During the subsequent inter-arrival interval I_2, some of these n packets may depart, while additional packets may arrive from the other $N-1$ users. As a result, when the next packet arrives in queue 2, it observes k packets in the system, where k jointly depends on the initial queue length n, the departures during I_2, and the random arrivals of the aggregated traffic.

To evaluate the sum $I_i + S_i$ where $i = \{1, 2\}$, we cannot rely on a convolution, but we need to consider the evolution of the queue state. For these reasons, we proceed by conditioning on the number of remaining packets k in the queue. Specifically, the system time follows an Erlang distribution, see formula (2.5), of order $k+1$, since the packet must wait for the completion of k service times plus its own. So we introduce two new quantities in our analysis:

- X_1: system time of a packet entering queue 1 conditioned on finding k packets in the queue

$$f_{X_1}(t) \sim \text{Erlang}(k+1, \mu_1) = \frac{\mu_1^{k+1} t^k e^{-\mu_1 t}}{k!},$$

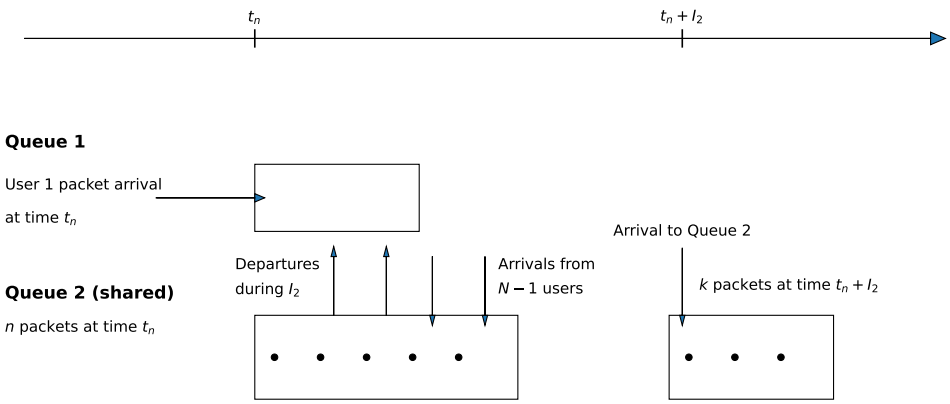

Figure 4.3: Illustration of packet dynamics in a two-queue system for a single user. Queue 1 is the private queue, while Queue 2 is shared among multiple users. The timeline shows the arrival of a packet to Queue 1 at time t_n and the evolution of Queue 2 during the inter-arrival interval I_2, including departures and arrivals from the other $N-1$ users. At time $t_n + I_2$, a new packet arrives to Queue 2, illustrating the change in queue occupancy from n to k packets.

- X_2: system time of a packet entering queue 2 conditioned on finding k packets in the queue

$$f_{X_2}(t) \sim \text{Erlang}(k+1, \mu_2) = \frac{\mu_2^{k+1} t^k e^{-\mu_2 t}}{k!}.$$

They both follow Erlang distributions with rates μ_1 and μ_2, respectively.

From this point onward, we introduce the following ancillary quantities:

- $a = (\mu_1 - \alpha\lambda)$,
- $b = (\mu_2 - N(1-\alpha)\lambda)$,
- $c = (1-\alpha)\lambda$,
- $d = \alpha\lambda$.

In this way, the expressions will become more compact and readable, simplifying the writing of probability calculations.

4.2.1 Packet Usefulness Analysis

We see that our process cannot be evaluated by treating the variables as independent, but we need to know the evolution of the queue state. In particular, the system time of the opposite queue depends on the number of packets n present when the packet we are analyzing arrives, on the inter-arrival time between packets and the k remaining packets in the queue. The value of k itself depends on n and the inter-arrival time. For simplicity, we perform the evaluation for queue 1 as the result for queue 2 is analogous.

The first step is therefore to compute the probability by conditioning it on n and weighting this by the stationary distribution ρ_2 as found in (2.3).

$$P(T_1 < I_2 + X_2) = \sum_{n=0}^{\infty} (1-\rho_2)\,\rho_2^n \cdot P(T_1 < I_2 + X_2 | N = n). \tag{4.2}$$

To evaluate $P(T_1 < I_2 + X_2 | N = n)$, we then consider all possible inter-arrival times, since X_2 depends on I_2 (but not vice versa).

$$P(T_1 < I_2 + X_2 | N = n) = \int_0^{\infty} c\, e^{-ct}\, P(T_1 < t + X_2 | N = n, I_2 = t)\, dt. \tag{4.3}$$

where ce^{-ct} is the inter-arrival pdf $f_{I_2}(t)$ defined previously. We now derive $P(T_1 < t + X_2|N = n, I_2 = t)$ conditioning on k, i.e. on the number of packets that the packet entering queue 2 finds in front of itself, given that n were there I_2 seconds earlier.

During the inter-arrival time $I_2 = t$, with $i = \{1, 2\}$, no additional packets from User 1 arrive, but som from the other users may. If we start with n packets in the queue, each will be delivered with rate μ_2 while the number of new packets that can arrive depends on the parameter $\lambda_2^\star = (N-1)(1-\alpha)\lambda$, which contains only the contributions from other users only. This is a birth-death process (see section 2.3) defined by the infinitesimal generator matrix:

$$Q = \begin{pmatrix} -\lambda_2^\star & \lambda_2^\star & 0 & 0 & \cdots \\ \mu & -(\lambda_2^\star + \mu_2) & \lambda_2^\star & 0 & \cdots \\ 0 & \mu & -(\lambda_2^\star + \mu_2) & \lambda_2^\star & \cdots \\ 0 & 0 & \mu & -(\lambda_2^\star + \mu_2) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The probability of having k remaining packets at time $I_2 = t$ is therefore:

$$p_k(t) = [\mathbf{p}(t)]_k = [\mathbf{p}(0)\, e^{Qt}]_k\,,$$

where $\mathbf{p}(0)$ represents the initial state distribution of the Markov process. In this case, the queue does not start empty but with n packets. Hence, the initial distribution is:

$$p_k(0) = \begin{cases} 1, & \text{if } k = n, \\ 0, & \text{if } k \neq n. \end{cases}$$

Once the distribution of k is known, it is possible to compute the conditional probability $P(T_1 < I_2 + X_2|K = k)$. This probability can be evaluated by integration:

$$\begin{aligned} P(T_1 < t + X_2 \mid K = k) &= \int_0^\infty P(T_1 < t + s)\, f_{X_2}(s)\, ds \\ &= \int_0^\infty (1 - e^{-a(t+s)})\, \frac{\mu_2^{k+1} s^k e^{-\mu_2 s}}{k!}\, ds \\ &= 1 - e^{-at} \left(\frac{\mu_2}{\mu_2 + a}\right)^{k+1}. \end{aligned} \tag{4.4}$$

Finally, by summing over all possible values of k, we obtain the exact conditional probability

$$P(T_1 < t + X_2|N = n, I_2 = t) = \sum_{k=0}^{\infty} p_k(t) \cdot P(T_1 < I_2 + X_2|K = k). \tag{4.5}$$

Combining (4.5) with (4.2) and (4.3), we find an exact formula for the probability that a packet is useful.

We also evaluate our analytical result with a Monte Carlo simulation, and, as shown in Figures 4.4 and 4.5, the simulation perfectly matches the theoretical predictions, confirming the validity of our model.

Figure 4.4 displays the probability of packet usefulness as a function of the splitting factor α with λ=1, N=2, μ_1=1 and μ_2=2. For low values of α, most packets are directed to queue 2, which becomes highly loaded. For this reason, the packet usefulness of queue 1 is greater than that of the shared queue; the few packets that arrive are more likely to be served immediately and to be useful. The two curves intersect at approximately α=0.2; beyond this point, the usefulness of queue 2 becomes higher than that of queue 1, while the usefulness of queue 1 decreases due to increased congestion. Choosing an appropriate α can balance usefulness between queues and traffic splitting strategies that could

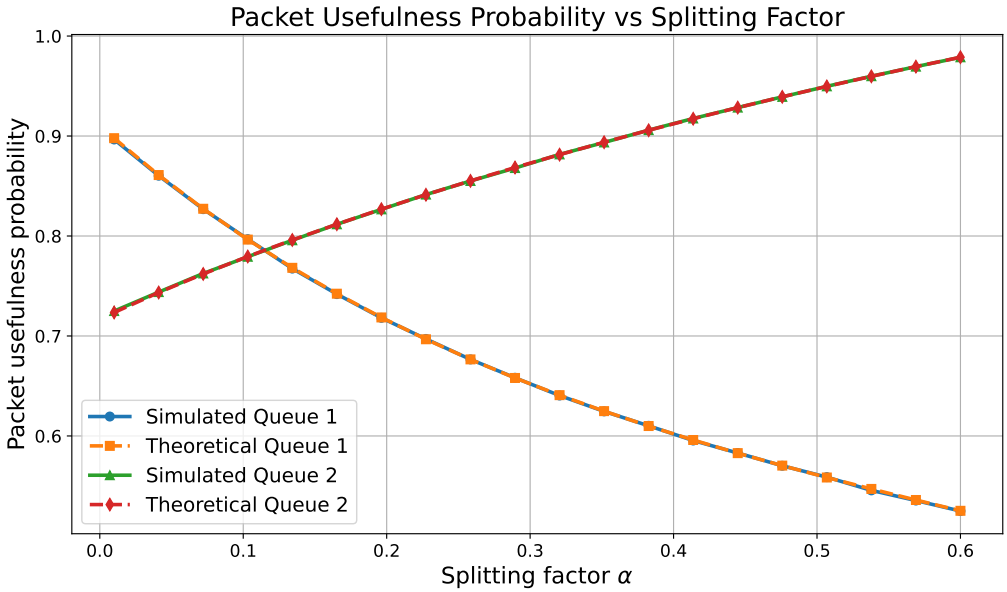

Figure 4.4: Packet usefulness probability as a function of the splitting factor α, with λ=1, N=2, μ_1=1 and μ_2=2.

account for current queue occupancy to reduce wasted service and improve effective AoI updates. Figure 4.5 shows the probability of packet usefulness as a function of the arrival rate λ. In this case, we set α=0.3, N=2, μ_1=1 and μ_2=2. For small values of λ, both queues are lightly loaded and the probability of being useful increases. As λ increases, the probability in the two queues decreases, reflecting increased congestion. Interestingly, the probability for queue 1 starts to increase again for $\lambda \gtrsim 0.9$. This behavior is likely due to the relatively low value of α, which results in fewer packets being assigned to queue 1. As queue 2 becomes heavily congested, packets in queue 1 are less likely to be preempted, leading to a slight increase in their probability of being useful.

Figure 4.6 highlights the difference in the usefulness of packets as a function of the total number of users. Considering α=0.3, λ=0.7, μ_1=2 and μ_2=3, the results are obtained using the theoretical expression validated previously. It can be observed that, with fewer users, fewer packets in the shared queue are preempted. This implies that it is not always convenient to direct all traffic to the private queue, since queue 2 may perform better when it is not heavily congested. In conclusion, these results provide insight into how queue parameters, such as splitting factor and arrival rate, influence packet usefulness. Understanding these dependencies could be valuable for designing systems that optimize information freshness and packet efficiency under different traffic conditions.

4.2.2 Shared Queue Packet Usefulness

The shared queue packet usefulness is a special case of the previous. In this case, we know that in the opposite queue there is only one user, so $\lambda_1^\star = 0$. This consideration brings a lot of simplification and allows us to find a closed-form expression. For completeness, we present the full derivation.

We fix an arrival of interest in queue 1 and denote I_1 its inter-arrival time. Let n be the number of packets present in queue 1 at the beginning of this interval. During the inter-arrival interval, no packets arrive in this queue, while service completions occur at rate μ_1. Our purpose is to compute the probability that a packet in the shared queue becomes *useless* because it is superseded by a newer packet from the private queue.

Given that k packets remain in queue 1 at time t, we can compute the probability as a function of k

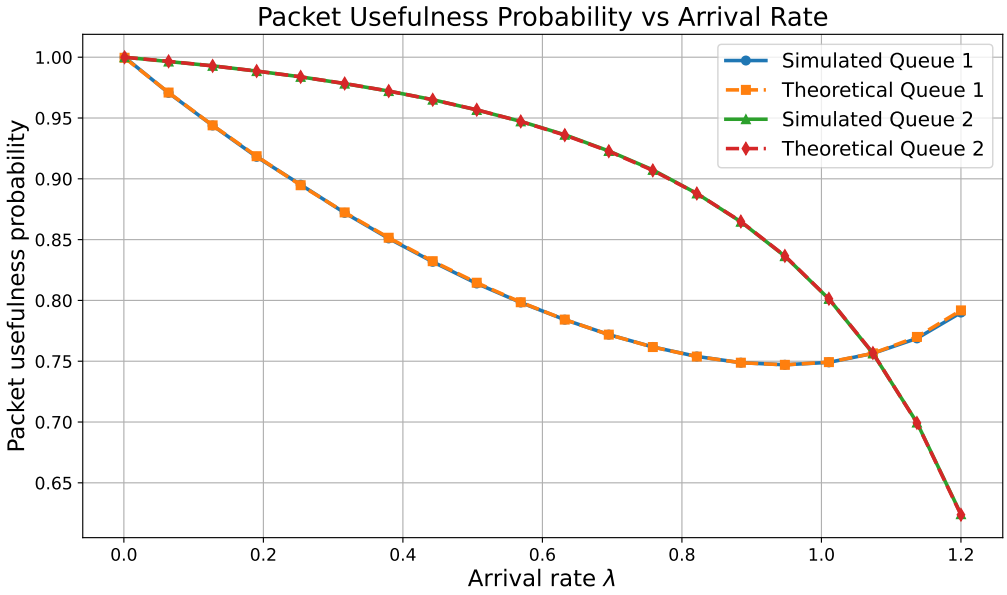

Figure 4.5: Packet usefulness probability as a function of the arrival rate λ, with α=0.3, N=2, μ_1=1 and μ_2=2.

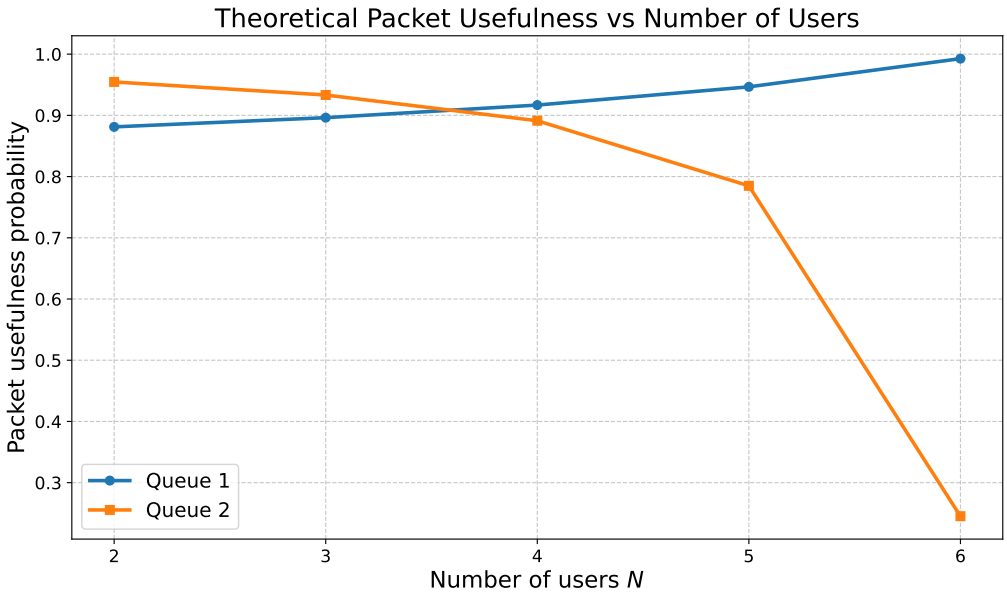

Figure 4.6: Packet usefulness probability as a function of the number of users N, with α=0.3, λ=0.7, μ_1=2 and μ_2=3.

as:

$$\begin{aligned} P(T_2 > t + X_1 \mid K = k) &= \int_0^\infty P(T_2 > t + s)\, f_{X_1}(s)\, ds \\ &= \int_0^\infty e^{-b(t+s)}\, \frac{\mu_1^{k+1} s^k e^{-\mu_1 s}}{k!}\, ds \\ &= e^{-bt} \left(\frac{\mu_1}{\mu_1 + b}\right)^{k+1}. \end{aligned}$$

After obtaining this expression, the question that we should ask is: how can we know the probability of having k remaining packets if we have n initial packets in the queue? The packets are served in an M/M/1 FCFS queue, so service completion follows a Poisson process with parameter $\mu_1 t$. The probability that exactly $n - k$ packets are served within time t is given by the Poisson distribution:

$$P_{\mu_1 t}(n - k) = \frac{(\mu_1 t)^{n-k} e^{-\mu_1 t}}{(n - k)!}.$$

However, we need to take in account that this formula describes a countable process, while in our system at most n packets can be delivered. So the conditional probability $P(k|n)$ follows a truncated Poisson process. It coincides with the Poisson probabilities until $n - 1$ packets are served, while the scenario with zero packets in the queue $k = 0$ corresponds to the cumulative probability:

$$1 - \sum_{k=0}^{N-1} P_{\text{Poisson}}(k; \mu_1 t).$$

Now that we evaluate $P(k|n)$, we can compute the probability as a function of n and t only, removing the parameter k by appropriately weighting over all its possible values:

$$\begin{aligned} \sum_{k=0}^{n} P(T_2 > t + X_1 \mid N = n, I_1 = t)\, P(k \mid n) &= \sum_{k=1}^{n} e^{-bt} \left(\frac{\mu_1}{\mu_1 + b}\right)^{k+1} P_{\mu_1 t}(k) \\ &\quad + e^{-bt} \left(\frac{\mu_1}{\mu_1 + b}\right) \left[1 - \sum_{k=0}^{n-1} P_{\mu_1 t}(k)\right] \\ &= f(n, t). \end{aligned}$$

For simplicity, we denote this probability $f(n, t)$. Now we just have to integrate over all the possible inter-arrival times:

$$P(T_2 > I_1 + X_1 | N = n) = \int_0^\infty d\, e^{-dt}\, f(n, t)\, dt,$$

where de^{-dt} is the pdf $f_{I_1}(t)$. Finally, we just need to sum over all possible values of n:

$$P(T_2 > I_1 + X_1) = \sum_{n=0}^{\infty} (1 - \rho_1)\, \rho_1^n \int_0^\infty d\, e^{-dT}\, f(n, t)\, dt.$$

where $\rho_1 = d/\mu_1$.

An important consideration at this point is that all the involved quantities are non-negative, since we are working with probabilities. Thus, applying Tonelli's theorem, we exchange the order of summation and integration. Finally, taking into account that the preceding calculation was performed for the probability of a packet being useless, we consider its complement to obtain the exact closed-form expression for the probability that a packet is useful:

$$P(T_2 < I_1 + T_1) = 1 - (1 - \rho_1)\, \frac{\mu_1}{(\mu_1 + b)^2} \left(\frac{d^2}{\mu_1 + b - d} + \frac{d(\mu_1 + b + d)}{b + d}\right).$$

If we limit our analysis to just one user, also for the private queue, we can use the same procedure. In this case, we only need to replace ρ_2 with ρ_1, a with b and c with d, thus obtaining:

$$P(T_1 < I_2 + T_2) = 1 - (1 - \rho_2)\, \frac{\mu_2}{(\mu_2 + a)^2} \left(\frac{c^2}{\mu_2 + a - c} + \frac{c(\mu_2 + a + c)}{a + c}\right).$$

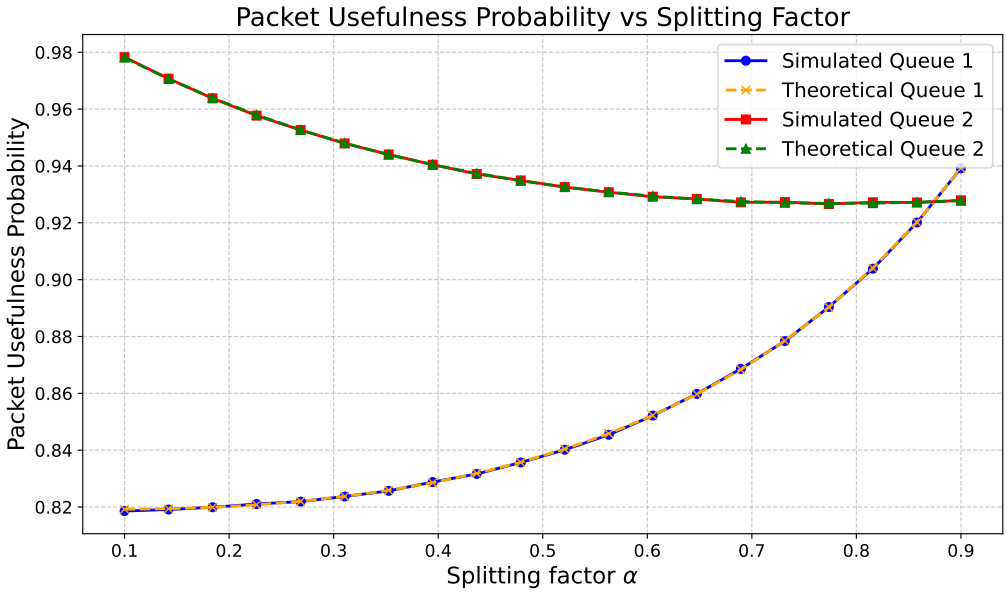

Figure 4.7: Packet usefulness probability as a function of the splitting factor α with one user and with λ=1, μ_1=2, μ_2=3

As we can see in Figure 4.7, when there is only one user splitting traffic between the two queues, the closed-form expression matches the simulation results. We can also notice that the shared queue has a packet usefulness higher than the private queue, confirming the importance of splitting the traffic instead of using only the private queue.

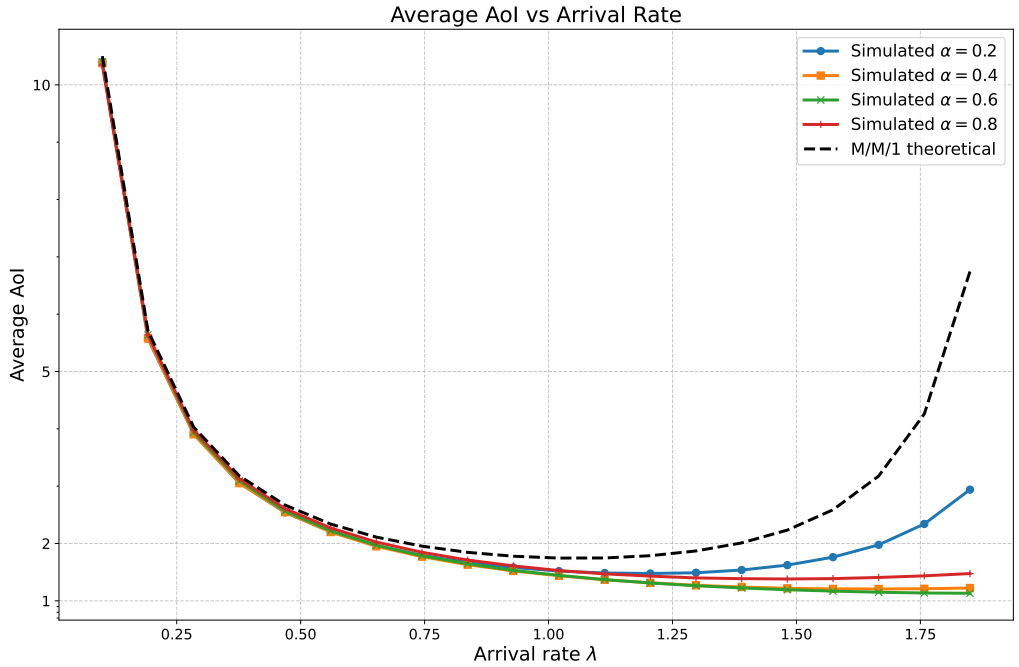

Figure 4.8: Average AoI versus arrival rate λ for different splitting factors $\alpha = 0.2, 0.4, 0.6, 0.8$, with $N = 2$, $\mu_1 = 2$, and $\mu_2 = 3$.

4.3 AoI Performance and Energy Trade-off

Obtaining a closed-form analytical expression for the average AoI in our multi-user, multi-queue system is highly challenging. Simplified formulas such as (2.2) can be used for particular queueing models (see Chapter 3.3), but our scenario is more complex.

For this reason, we study the AoI in the shared queue setting via Monte Carlo simulations. These simulations allow us to explore a wide range of parameter values and to gain insight into how the system behaves. They help us identify regions where our model exhibits improved performance in terms of information freshness.

Moreover, thanks to the notion of *packet usefulness probability*, we are also able to quantify how many packets update the AoI and how many, instead, are wasted due to preemption or obsolescence. This provides a more complete picture of the efficiency of the system, not only in terms of AoI but also in terms of resource usage.

To put our results in perspective, we compare the simulated AoI of our model with the well-known closed-form expression of the average AoI for a stationary M/M/1 queue with service rate μ_1, given in (2.7) [5]. Figure 4.8 shows how the average AoI changes as a function of the generation rate λ for an M/M/1 queue and for some values of the percentage of splitting α, considering $N = 2$, $\mu_1 = 2$ and $\mu_2 = 3$. We can see that in all cases the decision to send data only in the private queue is not efficient in terms of *freshness*. Moreover, we see that the average AoI decreases in α. However, at a certain point, it is no longer convenient to increase the value of α because there is a turning point in which the average AoI restarts to grow. In our graph, this is reached for α=0.6.

Based on the theoretical analysis of an M/M/1 queue with service rate $\mu_1 = 2$, the minimum average AoI is obtained at an arrival rate $\lambda_{\min}^{\mathrm{M/M/1}} = 1.064$, yielding $\Delta_{\min}^{\mathrm{M/M/1}} = 1.742$. This represents the optimal operating point in a single-server system where the freshness of information is maximized.

We then extended this analysis to a system in which the user splits its traffic between two queues with a routing probability $\alpha = 0.4$. By performing Monte Carlo simulations on the same range of arrival rates, we identified that, to achieve the same AoI as the M/M/1 minimum, $\Delta_{\alpha=0.4} = \Delta_{\min}^{\mathrm{M/M/1}}$, the corresponding arrival rate is $\lambda_{\alpha=0.4} = 0.761$. Interestingly, this configuration attains the same level of

timeliness while requiring a substantially lower effective arrival rate, approximately 28.5% less than in the single-server system.

These results suggest that introducing a controlled traffic split between queues allows the system to maintain comparable AoI performance while reducing the overall server load, highlighting the potential for more efficient resource utilization without compromising information freshness.

In conclusion, the multi-user multi-queue model shows a lot of improvements with respect to the stationary M/M/1 model. Using a hybrid GEO–LEO architecture, rather than relying on a single satellite, enables both improved information freshness and substantial reductions in energy expenditure. This shows that intelligent queueing and system design can fundamentally reshape the AoI and the energy trade-off.

5 Conclusions

This thesis investigated the behavior of the AoI under intermittent connectivity in multi-user, multi-queue systems, a scenario representative of NTNs. The study focused on understanding how system parameters, such as the service rate, the packet generation rate, the durations of the active and inactive phases, and the traffic splitting strategies, affect both information freshness and resource utilization.

In Chapter 3, we analyzed a system with deterministic active and inactive intervals and derived an exact closed-form expression for the AoI. This analytical result simplifies the evaluation of AoI and allows the identification of optimal operating points, as illustrated in Figures 3.12 and 3.13. These figures demonstrate the trade-off between timeliness and resource cost, enabling informed choices for service and generation rates. Figures 3.14 and 3.15 further highlight the impact of phase durations, showing that increasing the active-phase ratio improves information freshness, whereas longer inactive periods lead to higher AoI. These insights confirm that transmission strategies can be optimized even for moving satellites with intermittent line-of-sight connectivity.

Chapter 4 extended the analysis to multi-user, multi-queue systems and introduced analytical formulas for the packet usefulness, quantifying the fraction of updates that effectively contribute to reducing AoI. This provides a framework to optimize both information freshness and transmission efficiency under varying traffic conditions. Figure 4.6 illustrates that splitting the traffic between private and shared queues can be beneficial, particularly when queues are not heavily congested, as it increases packet usefulness. Simulation results, summarized in Figure 4.8, demonstrate that the multi-user, multi-queue system outperforms the single-server M/M/1 baseline. The analysis also shows that an optimal traffic-splitting factor α exists and that a comparable AoI can be achieved with a lower effective arrival rate, reducing the number of transmissions and the associated energy consumption.

These findings have significant implications for satellite communication systems. Using hybrid GEO–LEO architectures, rather than relying on a single satellite, can enhance the freshness of information while reducing energy expenditure. Moreover, intelligent traffic management through queue splitting and optimized arrival rates enables efficient resource utilization without compromising system performance.

While the models developed in this work capture key aspects of AoI under intermittent connectivity and highlight the role of packet usefulness in multi-user, multi-queue systems, an exact analytical evaluation of AoI for the full multi-user, multi-queue scenario remains an open challenge. Future research could focus on extending the analytical framework to address this problem, as well as exploring adaptive routing and scheduling strategies to further improve system efficiency.

5.1 Future Works

While this thesis has provided insights into AoI dynamics in intermittent connectivity scenarios and multi-user, multi-queue systems, several promising directions remain open for future investigation. One natural extension is the analysis of multi-hop or mesh satellite networks. Building on studies such as [30], the current model could be generalized to account for chains of intermediate nodes, each with its own queueing behavior and service characteristics. This extension would enable the evaluation of AoI across a chain of relays. It would capture the compounding effects of delays, packet loss, and queue interactions, providing a more realistic representation of NTNs.

Another interesting avenue is the incorporation of heterogeneous users and multiple queue disciplines,

including FCFS, LCFS, and replacement policies [39]. The model could be extended to scenarios with several classes of users generating traffic with different patterns and priorities. Studying how AoI evolves under multi-priority or heterogeneous traffic conditions would provide insights into optimal queue management and scheduling.

Energy-aware optimization represents a third critical direction. Modern NTN and IoT scenarios frequently involve battery-powered devices or satellites with limited energy resources. Future work could incorporate energy harvesting and power constraints, balancing information freshness with energy expenditure [39] [40]. This would allow the design of energy-efficient transmission policies that dynamically adjust traffic splitting or scheduling based on both AoI and available energy, reflecting realistic operational limitations.

Finally, a broader perspective involves bridging the gap between theoretical models and real systems. Real-world networks introduce challenges such as variable channels, fading, mobility, and packet loss [41]. Future research could leverage the analytical foundations developed in this thesis to design models that account for these stochastic phenomena, potentially combining analytical, simulation-based, and machine learning approaches to handle complex and dynamic scenarios. This would enable the deployment of AoI-aware strategies in realistic LEO–GEO satellite systems, edge computing networks and IoT applications, providing a comprehensive understanding of the trade-offs between information freshness, latency and energy consumption.

6
Bibliography

[1] David S. Alberts and Daniel S. Papp, eds. *The Information Age: An Anthology on Its Impact and Consequences*. National Defense University Press, 1997.

[2] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[3] Daniel J. Costello and G. David Forney. "Channel coding: The road to channel capacity". In: *Proceedings of the IEEE* 95.6 (2007), pp. 1150–1177.

[4] Sanjit Kaul, Roy Yates, and Marco Gruteser. "Real-time status: How often should one update?" In: *2012 Proceedings IEEE INFOCOM*. 2012, pp. 2731–2735.

[5] Roy D. Yates et al. "Age of Information: An Introduction and Survey". In: *CoRR* abs/2007.08564 (2020). arXiv: 2007.08564.

[6] Yoshiaki Inoue et al. "A General Formula for the Stationary Distribution of the Age of Information and Its Application to Single-Server Queues". In: *IEEE Transactions on Information Theory* 65.12 (2019), pp. 8305–8324.

[7] Sanjit Kaul et al. "Minimizing age of information in vehicular networks". In: *2011 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*. 2011, pp. 350–358.

[8] Sanjit Kaul, Roy Yates, and Marco Gruteser. "On Piggybacking in Vehicular Networks". In: *2011 IEEE Global Telecommunications Conference - GLOBECOM 2011*. 2011, pp. 1–5.

[9] Maice Costa, Marian Codreanu, and Anthony Ephremides. "On the Age of Information in Status Update Systems With Packet Management". In: *IEEE Transactions on Information Theory* 62.4 (2016), pp. 1897–1910.

[10] Maice Costa, Marian Codreanu, and Anthony Ephremides. "Age of information with packet management". In: *2014 IEEE International Symposium on Information Theory*. 2014, pp. 1583–1587.

[11] Mustafa Emara, Hesham ElSawy, and Gerhard Bauch. "A Spatiotemporal Framework for Information Freshness in IoT Uplink Networks". In: *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*. 2020, pp. 1–6.

[12] Juan Mena and Felipe Núñez. "Age of information in IoT-based networked control systems: A MAC perspective". In: *Automatica* 147 (Jan. 2023), p. 110652.

[13] Arunabh Srivastava, Abhishek Sinha, and Krishna Jagannathan. "On Minimizing the Maximum Age-of-Information For Wireless Erasure Channels". In: *2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*. 2019, pp. 1–6.

[14] Leonardo Badia, Andrea Zanella, and Michele Zorzi. "A Game of Ages for Slotted ALOHA With Capture". In: *IEEE Transactions on Mobile Computing* 23.5 (2024), pp. 4878–4889.

[15] Haoyuan Pan et al. "Age of Information With Collision-Resolution Random Access". In: *IEEE Transactions on Vehicular Technology* 71.10 (2022), pp. 11295–11300.

[16] Leonardo Badia and Andrea Munari. "Satellite Intermittent Connectivity and Its Impact on Age of Information for Finite Horizon Scheduling". In: *2025 12th Advanced Satellite Multimedia Systems Conference and the 18th Signal Processing for Space Communications Workshop (ASMS/SPSC)*. 2025, pp. 1–8.

[17] Lucrezia Rossi and Leonardo Badia. "Cyclic Intermittent Connectivity of Industrial Sensors Impacting Information Freshness". In: *2025 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0 & IoT)*. 2025, pp. 422–427.

[18] Mohammad Moltafet, Markus Leinonen, and Marian Codreanu. "On the Age of Information in Multi-Source Queueing Models". In: *IEEE Transactions on Communications* 68.8 (2020), pp. 5003–5017.

[19] Sanjit K. Kaul and Roy D. Yates. "Timely Updates By Multiple Sources: The M/M/1 Queue Revisited". In: *2020 54th Annual Conference on Information Sciences and Systems (CISS)*. 2020, pp. 1–6.

[20] Anhad Bhati, Sibi Raj B Pillai, and Rahul Vaze. "On the Age of Information of a Queuing System with Heterogeneous Servers". In: *2021 National Conference on Communications (NCC)*. 2021, pp. 1–6.

[21] Marco Giordani and Michele Zorzi. "Non-Terrestrial Networks in the 6G Era: Challenges and Opportunities". In: *IEEE Network* 35.2 (2021), pp. 244–251.

[22] Leonardo Badia and Maria Scalabrin. "Stochastic Analysis of Delay Statistics for Intermittently Connected Vehicular Networks". In: *European Wireless 2014; 20th European Wireless Conference*. 2014, pp. 1–6.

[23] Yuanzhi Ni, Lin Cai, and Yuming Bo. "Vehicular beacon broadcast scheduling based on age of information (AoI)". In: *China Communications* 15.7 (2018), pp. 67–76.

[24] Leonardo Badia et al. "Routing schemes in heterogeneous wireless networks based on access advertisement and backward utilities for QoS support [Quality of Service based Routing Algorithms for Heterogeneous Networks]". In: *IEEE Communications Magazine* 45 (2007), pp. 67–73.

[25] Lawler G.F. *Introduction to Stochastic Processes (2nd ed.)* Chapman and Hall/CRC, 2006.

[26] 3GPP. *5G Non-Terrestrial Networks*. 3GPP news release. Jan. 25, 2024. URL: `https://www.3gpp.org/news-events/3gpp-news/5g-ntn`.

[27] Oltjon Kodheli et al. "Satellite Communications in the New Space Era: A Survey and Future Challenges". In: *IEEE Communications Surveys & Tutorials* 23.1 (2021), pp. 70–109.

[28] Federica Rinaldi et al. "Non-Terrestrial Networks in 5G & Beyond: A Survey". In: *IEEE Access* 8 (2020), pp. 165178–165200.

[29] Andrew S. Tanenbaum and David J. Wetherall. *Computer Networks*. 5th. Upper Saddle River, NJ: Pearson, 2010.

[30] Federico Chiariotti et al. "Information Freshness of Updates Sent over LEO Satellite Multi-Hop Networks". In: *CoRR* abs/2007.05449 (2020).

[31] Beatriz Soret, Sucheta Ravikanti, and Petar Popovski. "Latency and timeliness in multi-hop satellite networks". In: *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*. 2020, pp. 1–6.

[32] Georgios Bouloukakis et al. "Simulation-based Queueing Models for Performance Analysis of IoT Applications". In: *2018 11th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP)*. 2018, pp. 1–5.

[33] Vincenzo Mancuso et al. "Effectiveness of distributed stateless network server selection under strict latency constraints". In: *Computer Networks* 251 (June 2024), p. 110558.

[34] M. Mitzenmacher. "The power of two choices in randomized load balancing". In: *IEEE Transactions on Parallel and Distributed Systems* 12.10 (2001), pp. 1094–1104.

[35] Weisong Shi et al. "Edge Computing: Vision and Challenges". In: *IEEE Internet of Things Journal* 3.5 (2016), pp. 637–646.

[36] Mung Chiang and Tao Zhang. "Fog and IoT: An Overview of Research Opportunities". In: *IEEE Internet of Things Journal* 3.6 (2016), pp. 854–864.

[37] Xiongwei Wu et al. "Deep Reinforcement Learning for loT Networks: Age of Information and Energy Cost Tradeoff". In: *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*. 2020, pp. 1–6.

[38] Jie Gong, Xiang Chen, and Xiao Ma. "Energy-Age Tradeoff in Status Update Communication Systems with Retransmission". In: *2018 IEEE Global Communications Conference (GLOBECOM)*. 2018, pp. 1–6.

[39] Mehrdad Salimnejad and Nikolaos Pappas. "On the Age of Information in a Two-User Multiple Access Setup". In: *Entropy* 24.4 (2022), p. 542.

[40] Huixiang Zhao et al. “AoI-and-energy tradeoff scheduling for multi-UAV-enabled data acquisition in Wireless Sensor Networks”. In: 178 (July 2025), p. 103985.

[41] Elena Zhbankova et al. “The Age of Information in Wireless Cellular Systems: Gaps, Open Problems, and Research Challenges”. In: *Sensors* 23.19 (2023), p. 8238.