Analysis and Optimization of Age of Information for Real-Time Communications

Zhifeng Tang

February 2024

A Thesis Submitted for the Degree of
Doctor of Philosophy
of The Australian National University

School of Engineering
College of Engineering, Computing and Cybernetics
The Australian National University

Declaration

The contents of this thesis are the results of original research and have not been submitted for a higher degree to any other university or institution.

The research work presented in this thesis has been performed jointly with Assoc. Prof. Nan Yang (The Australian National University), Assoc. Prof. Zhuo Sun (Northwestern Polytechnical University, China), Assoc. Prof. Xiangyun Zhou (The Australian National University), Prof. Parastoo Sadeghi (University of New South Wales), and Prof. Jemin Lee (Sungkyunkwan University, Republic of Korea). The substantial majority of this work was my own.

Much of the work in this thesis has been published or has been submitted for publication as journal papers or conference proceedings.

Zhifeng Tang
School of Engineering
College of Engineering, Computing and Cybernetics
The Australian National University
Canberra, ACT 2601
Australia

Acknowledgments

I would like to first express my deepest gratitude to my principal supervisor Assoc. Prof. Nan Yang for giving me the opportunity to pursue the doctoral degree under his supervision. I really appreciate his unwavering support, guidance, and encouragement throughout my PhD journey and his expertise, valuable insights, and constructive feedback have been instrumental in shaping the direction and quality of my research. My special thanks to Assoc. Prof. Xiangyun Zhou and Prof. Parastoo Sadeghi for being part of my supervisory panel. I really appreciate their effort and time for evaluating my annual progress and providing guidance regarding to my PhD study.

I would also like to express my sincere appreciation to Assoc. Prof. Zhuo Sun at the Northwestern Polytechnical University (China) for her guidance, help, and valuable inputs to our collaborated research works. Many thanks to my collaborators Prof. Jemin Lee at the Sungkyunkwan University (Republic of Korea) for her constructive comments and helpful discussions to our research works.

I am grateful to the Australian National University (ANU) for providing me the PhD scholarship and stipend for my PhD study. Big thanks to the support from the Australian Research Council (ARC) Discovery Project (DP180104062): “Ultra-Reliable and Low-Latency Mission Critical Communications” for my PhD study in ANU and conference attendance.

I am immensely thankful to my colleagues in the Communications Group in the School of Engineering at ANU for creating a friendly and relaxing study environment. To all the colleagues both in past and present that I met in the office, especially Prof. Salman Durrani, Prof. Thushara D. Abhayapala, Dr. Chunhui Li, Dr. Muhammed Akram Mohamed Shafie, Dr. Xinyu Huang, Dr. Chiu Chun Chan, Dr. Yuting Fang, Dr. Yucheng Liu, Dr. Sheeraz Alvi, Dr. Xiaohui Zhou, Dr. Xiaolun Jia, Dr. Milad Tatar Mamaghani, Hang Yuan, Simin Xu, Sahar Idrees, Tayyaba Ilyas, Liangling Lai, Mohammad Amin Zarrabian, Nilupuli Senadhira, Qingyang Zhung, and Saud Khan.

I would like to express my deepest gratitude and love to my parents Chengwan Tang and Yonghong Jiang, my parents-in-law Zhujiang Huang and Bihong Wang, and my wife Xuanyi Huang for their unconditional love, unlimited support, and continuing encouragement. I am sincerely grateful to my wife for supporting and accompanying me during the ups and downs moments along these years. This thesis is dedicated to my family, especially to my wife.

Abstract

In order to meet the increasing demand for wireless communication, the industry and academia in the telecommunications sector have shifted their interests towards exploring the next generation communication technologies, known as the sixth generation (6G) communications. In the envisioned 6G communications, many applications and services rely on real-time communications where timely information delivery is essential. To characterize the freshness of delivered information, an emerging performance metric named age of information (AoI) has been introduced, which is defined as the elapsed time since the last successfully received packet was generated by the transmitter.

Motivated by the importance of AoI, this thesis aims to analyze and optimize the AoI performance in wireless networks. Specifically, this thesis investigates and addresses the following issues: (i) Transmission scheduling policy design in multi-user systems (Chapter 2), (ii) Offloading strategy design in multi-user mobile edge computing (MEC) systems (Chapter 3), (iii) Analysis of AoI performance for different encoding schemes and non-linear AoI performance in short packet communication systems (Chapter 4), and (iv) Optimal transmission selection in short packet communications (Chapter 5). The contributions made in this thesis are summarized as follows:

First, we design scheduling policies of multi-user systems to optimize the AoI performance in Chapter 2. In this chapter, we design two different scheduling policies, where the first one is a decentralized uplink packet arrival-independent scheduling policy and the second one is a centralized downlink packet arrival-dependent scheduling policy. In particular, we first propose a repeated random access (RA) based scheduling policy, where each user equipment (UE) generates its packet and then randomly selects one channel from multiple orthogonal frequency channels to perform repeat transmission of its packet across multiple time slots. We derive closed-form expressions for the average AoI achieved by the proposed scheduling policy for three packet management policies. The simulation results corroborate our analysis and demonstrate that the proposed scheduling policy can achieve a lower average AoI than the round robin (RR) policy. We then propose a Whittle index based scheduling policy to minimize the average general AoI penalty in a downlink multi-user system with stochastic packet generation patterns and unreliable transmission links. We formulate the transmission scheduling problem as an average cost constrained Markov decision process (CMDP) problem. Through introducing the service charge, we derive the approximated expression for the Whittle index, based on which we design the scheduling policy. Using numerical results, we demonstrate the performance gain of our designed scheduling policy compared to the existing policies.

Second, we investigate the offloading strategy of a multi-user MEC system to achieve the optimal AoI performance in Chapter 3. We analyze the average AoI and the average peak AoI (PAoI) of a multi-user MEC system. In this MEC system, we derive closed-form expressions for the average AoI and average PAoI for three computing schemes, namely, the local computing scheme, the edge computing scheme, and the partial computing scheme. To reduce the complexity of the average AoI expression, we derive upper and lower bounds on the average AoI, which allow us to explicitly examine the dependence of the optimal offloading decision on the MEC system parameters. Aided by simulation results, we verify our analysis and illustrate the impact of system parameters on the AoI performance.

Third, we investigate the AoI performance of short packet communication systems in Chapter 4. In this chapter, we first analyze the impact of different encoding schemes on the AoI performance in a short packet communication system, where a transmitter generates packets based on the status updates collected from multiple sensors and transmits the packets to a receiver. In this system, we consider two encoding schemes, namely, the joint encoding scheme and the distributed encoding scheme. We derive closed-form expressions for the average AoI achieved by these two encoding schemes and compare their performance. Simulation results show that the distributed encoding scheme is more appropriate for systems with a relatively large number of sensors, compared with the joint encoding scheme. We then analyze the non-linear AoI performance in a short packet communication system, where a transmitter generates packets based on status updates and transmits the packets to a receiver. Specifically, we investigate three packet management strategies, namely, the non-preemption with no buffer strategy, the non-preemption with one buffer strategy, and the preemption strategy. To characterize the level of the receiver's dissatisfaction on outdated data, we incorporate a generalized α-β AoI penalty function into the analysis and derive closed-form expressions for the average AoI penalty achieved by three packet management strategies. Aided by simulation results, we verify our analysis and investigate how the value of α affects the system transmission reliability.

Finally, we analytically decide whether the broadcast transmission scheme or the unicast transmission scheme achieves the optimal AoI performance of a multi-user system in Chapter 5. In this system, we consider two transmission schemes, i.e., the broadcast transmission scheme and the unicast transmission scheme. For both transmission schemes, we examine three packet management strategies, namely the non-preemption strategy, the preemption in buffer strategy, and the preemption in serving strategy. We derive closed-form expressions for the average AoI achieved by two transmission schemes with three packet management strategies. Based on them, we compare the AoI performance of two transmission schemes in two systems, namely, the remote control system and the dynamic system. Aided by simulation results, we verify our analysis and investigate the impact of system parameters on transmission scheme selection to optimize the average AoI.

Contents

List of Publications

Journal Articles

J1. **Z. Tang**, Z. Sun, N. Yang, and X. Zhou, "Whittle index-based scheduling policy for minimizing the cost of age of information," *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 54-58, Jan. 2022. (Chapter 2)

J2. **Z. Tang**, Z. Sun, N. Yang, and X. Zhou, "Age of information of multi-user mobile edge computing systems," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1600-1614, Jul. 2023. (Chapter 3)

J3. **Z. Tang**, N. Yang, P. Sadeghi, and X. Zhou, "Age of information in downlink systems: Broadcast or unicast transmission?" *IEEE J. Select. Areas Commun.*, vol. 41, no. 7, pp. 2057-2070, Jul. 2023. (Chapter 5)

Conference Papers

C1. **Z. Tang**, Z. Sun, N. Yang, and X. Zhou, "Age of information of multi-source systems with packet management," in *Proc. IEEE Int. Conf. Commun. Workshop*, Dublin, Ireland, Jun. 2020, pp. 1-6. (Chapter 2)

C2. **Z. Tang**, Z. Sun, N. Yang, and X. Zhou, "Age of information analysis of multi-user mobile edge computing systems," in *Proc. IEEE Global Commun. Conf.*, Madrid, Spain, Dec. 2021, pp. 1-6. (Chapter 3)

C3. **Z. Tang**, N. Yang, P. Sadeghi, and X. Zhou, "The age of information of short-packet communications: Joint or distributed encoding?" in *Proc. IEEE Int. Conf. Commun.*, Seoul, Republic of Korea, May 2022, pp. 2175-2180. (Chapter 4)

C4. **Z. Tang**, N. Yang, X. Zhou, and J. Lee, "Age of information penalty of short-packet communications with packet management," in *Proc. IEEE Int. Conf. Commun.*, Rome, Italy, May 2023, pp. 1-6. (Chapter 4)

C5. **Z. Tang**, N. Yang, P. Sadeghi, and X. Zhou, "Broadcast versus distributed short-packet transmission: An age of information perspective," in *Proc. IEEE Int. Conf. Commun.*, Rome, Italy, May 2023, pp. 1-6. (Chapter 5)

List of Acronyms

2D	two-dimensional
3GPP	3rd Generation Partnership Project
5G	fifth generation
6G	sixth generation
ACK	acknowledgment
AI	artificial intelligent
AMC	adaptive modulation and coding
AoI	age of information
AoII	age of incorrect information
ARQ	automatic repeat request
AWGN	additive white Gaussian noise
BRNP	broadcast transmission with non-preemption
BRPS	broadcast transmission with preemption in serving
BS	base station
CDF	cumulative distribution function
CMDP	constrained Markov decision process
c.u.	channel use
eMBB	enhanced Mobile Broadband
FCFS	first-come-first-serve
HARQ	hybrid automatic repeat request
IoT	Internet of Things
ITU	International Telecommunication Union

LCFS	last-come-first-serve
LoS	line-of-sight
MAC	media access control
MEC	mobile edge computing
mMTC	massive machine-type Communications
ms	millisecond
MTC	machine type communication
NOMA	non-orthogonal multiple access
NP	non-deterministic polynomial-time
NPNB	non-preemption with no buffer
NPOB	non-preemption with one buffer
PAoI	peak age of information
PDF	probability density function
PGF	probability generating function
PMF	probability mass function
PPP	Poison point process
RA	random access
RR	round robin
SNR	signal-to-noise ratio
TB	transmission block
UAV	unmanned aerial vehicle
UE	user equipment
UNINP	unicast transmission with non-preemption
UNIPB	unicast transmission with preemption in buffer
UNIPS	unicast transmission with preemption in serving
URLLC	ultra-reliable and low latency communications

USRP	software-defined radio platforms
w.r.t.	with respect to

Notations

Δ	the average AoI
$\Delta_n(t)$	the AoI of the n-th UE at time t
Ω	the average PAoI
$\mathbb{E}[\cdot]$	the expectation operator
$\mathbb{1}(\cdot)$	indicator function
$\Pr(\cdot)$	probability mass function
$f(\cdot)$	probability density function
$f(\cdot\|\cdot)$	conditional probability density function
$\log_2(\cdot)$	logarithm to base 2
$Q(\cdot)$	Q-function

List of Figures

Chapter 1

Introduction

1.1 Background

The fifth generation (5G) communication markets have experienced significant revenue growth over the last few years, due to the staggering number of over one billion 5G subscriptions worldwide, reflecting the increasing demand of 5G services [1]. By the end of 2028, 5G subscriptions are projected to reach a global total of 4.6 billion. To address the ever-increasing demand for communication, the entire telecommunications sector, including both industry and academia, have shifted their focus towards exploring the next generation communication technologies, i.e., the sixth generation (6G) communications. The envisioned 6G communication systems aim to build upon the foundation laid by 5G and address the emerging demands of our ever-connected society. To further improve the cellular communication performance, the International Telecommunication Union (ITU) has proposed the framework of the envisioned initial 6G communications based on the 5G communications and plans to finish the 6G standardization process before 2030 [2]. The envisioned 6G communications will lead to a revolutionary leap forward in terms of coverage, data rates, network reliability, latency, massive connectivity, security, and energy efficiency. In 6G communications, the main usage scenarios are extended from the original three scenarios of 5G communications, i.e., enhanced Mobile Broadband (eMBB), ultra-reliable and low latency communications (URLLC), and massive machine-type communications (mMTC), to the six usage scenarios [3], as follows

- **Immersive communication**: This scenario is the enhancement of the eMBB, which aims to deliver enhanced accessibility to a wide range of multimedia content, services, and data, ensuring a seamless user experience.

- **Massive communication**: This scenario is the enhancement of the mMTC, which supports a massive number of low-power and low-complexity Internet of Things (IoT) devices.

- **Hyper reliable and low latency communication**: This scenario is the enhancement of the URLLC, which provides real-time communication services with extremely low latency and high reliability.

- **Ubiquitous connectivity**: This scenario is the first new usage scenario, which ensures that communication networks cover remote and underserved areas with reliable connectivity.

- **Integrated artificial intelligence (AI) and communication**: This scenario is the second new usage scenario, which aims to create an efficient, adaptive, and intelligent network to deliver a highly customized and context-aware user experience.

- **Integrated sensing and communication**: This scenario is the third new usage scenario, which aims to support real-time data acquisition from diverse sensors and create an interconnected and data-driven network.

In 6G communications, many applications and services rely on real-time communications, where timely information delivery is essential. Real-time communications support the capability of 6G communication systems to facilitate instantaneous and ultra-responsive data exchange between connected devices and systems. Particularly, timely wireless transmission is a crucial/vital requirement for realizing accurate monitoring and control in real-time communication applications [4, 5]. In such applications, maintaining the information freshness is critical, since the outdated information may result in the system failure and safety hazards. Therefore, real-time communications are key enablers to support industries which require instantaneous data transmission to ensure efficient operations, safety, and decision-making, such as aviation, maritime, emergency response, and financial services. Moreover, real-time communications are key pillars underpinning critical situations, such as disaster response or military operations, where timely information can save lives and prevent deaths. Furthermore, timely information delivery is vital for effective remote collaboration and connectivity. For example, in distributed work environment, real-time communication tools and access to shared information enable seamless collaboration among sensors and/or robots, regardless of their geographical locations, ensuring them to stay connected and work together efficiently.

1.2 Timeliness Performance Metric

When assessing the time-relevant performance of a communication system, two commonly utilized metrics are latency and status update frequency. Latency refers to the time delay between the generation of information from the source to its reception at the destination. Thus, minimizing latency is crucial for timely information delivery, as high latency may adversely affect applications such as video streaming, voice communication, online gaming, and real-time monitoring. Status update frequency, on the other hand, refers to how often the system updates its status. High status update frequency is crucial for real-time applications to ensure timely information dissemination, monitoring, and control.

The real-time applications require both low latency and high status update frequency transmission, while low latency reduces the delay in transmitting information and high status update frequency enables the system to provide more up-to-date information. However, there may exist a trade-off between these two performance metrics [6], where high status update frequency may lead to a large queuing delay which in turn increases latency, while decreasing latency may require a low status update frequency, resulting the lack of updates. Therefore, balancing both metrics is crucial for achieving the optimal performance in terms of timely and efficient information dissemination. In other words, to accurately measure and quantify the timeliness of information in real-time applications, it is essential to employ a suitable performance metric, which jointly considers latency and status update frequency.

1.2.1 Age of Information

In order to fully characterize the freshness of delivered information, the concept of age of information (AoI) was introduced as a new performance metric in [7], which is utilized to quantify the freshness of a communication system from the perspective of the receiver. Specifically, AoI is defined as the elapsed time since the last successfully received packet was generated by the transmitter, which is a time metric that characterizes the latency in status update systems and applications while considering the throughput of the system.

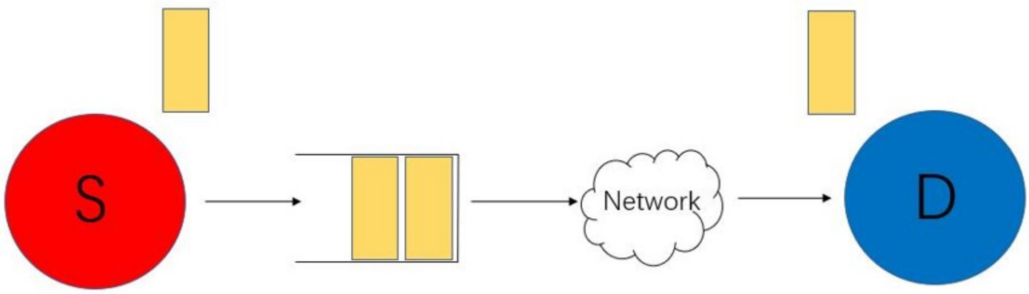


Figure 1.1: A source transmits state update packets to a destination through the wireless communication network.

A simple state update model is shown in Fig. 1.1, where a source S generates state update packets and transmits them to the destination D via a wireless communication network. In this system, let us denote $V(t)$ as the generation time of the last successfully detected packet at D and this introduces the AoI process $\Delta(t)$ shown in Fig. 1.2, which is characterized as

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

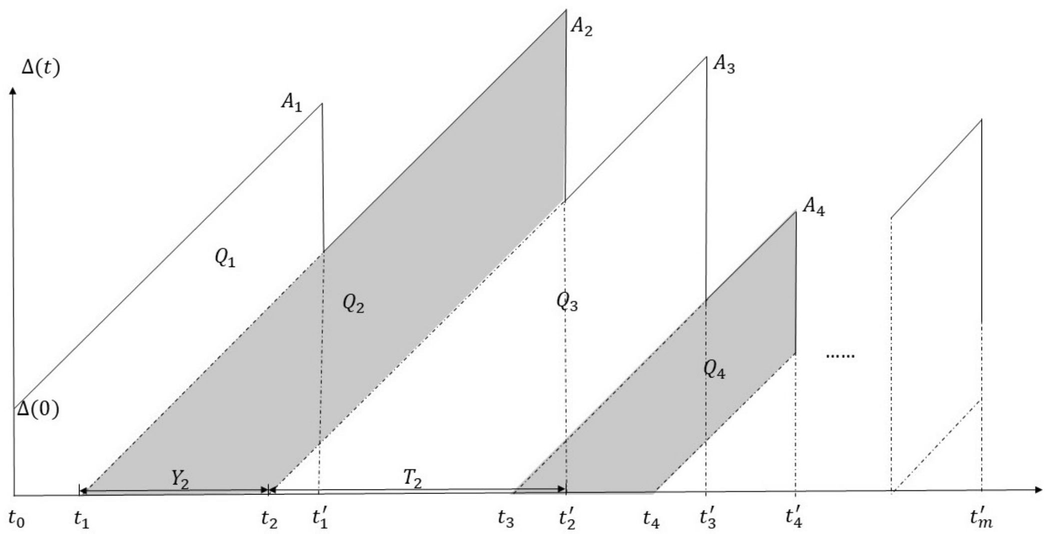


Figure 1.2: An example of the AoI evolution in time.

In Fig. 1.2, these transmitted packets are generated at times t_1, t_2, $\ldots$ and received at the time t_1', t_2', $\ldots$. When the jth packet is successfully delivered to D, $V(t)$ is updated as t_j and thus the AoI is reset to the delay that the packet experienced in the network, i.e., $\Delta(t_j') = t_j' - t_j$. In the absence of a newer update, the age process $\Delta(t)$ grows at unit rate, resulting in the characteristic sawtooth pattern of the AoI process.

1.2.2 Time Average AoI

Given an AoI process $\Delta(t)$, the average AoI can be calculated using a sample average. In particular, the time average AoI is given as

$$\Delta_\tau = \frac{1}{\tau}\int_0^\tau \Delta(t)\mathrm{d}t, \tag{1.2}$$

where $(0,\tau)$ is the interval of observation and the ensemble average AoI can be obtained by adopting $\tau \to \infty$. Considering that D receives the mth packet at the end of this time interval, i.e., $\tau = t_m'$, the integral part in (1.2), i.e., $\int_0^\tau \Delta(t)\mathrm{d}t$, is the sum of disjoint areas Q_j, $j = \{1,2,\cdots,m\}$, given by

$$\int_0^\tau \Delta(t)\mathrm{d}t = \sum_{j=1}^{m} Q_j + \frac{T_m^2}{2} = Q_1 + \frac{T_m^2}{2} + \sum_{j=2}^{m} Q_j. \tag{1.3}$$

From Fig. 1.2, the area Q_1 is a polygon and Q_j is an isosceles trapezoid for $j \geq 2$. Let denote $Y_j = t_j - t_{j-1}$ as the interarrival time, which is the time interval from the generation time of $j-1$th packet to the generation time of jth packet, and $T_j = t_j' - t_j$ as the latency, which is the time interval from the generation time of jth packet to the time that jth packet is successfully delivered to D. Hence, Q_j can be derived from two isosceles triangles, i.e.,

$$Q_j = \frac{1}{2}(Y_j + T_j)^2 - \frac{1}{2}T_j^2 = \frac{Y_j^2}{2} + Y_j T_j. \tag{1.4}$$

From (1.3), when $\tau \to \infty$, the impact of Q_1 and T_m^2 on the average AoI is negligible, i.e., $\lim_{\tau\to\infty} \frac{2Q_1 + T_m^2}{2\tau} = 0$, because Q_1 and T_m^2 are finite. When Y_j and T_j are stationary ergodic processes, the average AoI is given as

$$\Delta = \frac{\lim\limits_{m\to\infty} \frac{1}{m-1} \sum\limits_{j=2}^{m} Q_j}{\mathbb{E}[Y_j]} = \frac{\mathbb{E}[Q_j]}{\mathbb{E}[Y_j]} = \frac{\mathbb{E}[Y_j^2] + 2\mathbb{E}[Y_j T_j]}{2\mathbb{E}[Y_j]}. \tag{1.5}$$

The average AoI analysis in (1.5) can be applied to a wide range of service systems and it does not make any particular assumption on the traffic pattern and the queue discipline. It is noted that a substantial interarrival time Y_j may lead to a small queue size, resulting in a low waiting time and thus a low latency T_j. However, due to the correlation between Y_j and T_j, the evaluation of $\mathbb{E}[Y_j T_j]$ may be challenging, making the average AoI analysis difficult.

The average AoI is suitable for applications where maintaining up-to-date information is critical, such as real-time monitoring, surveillance systems, and distributed sensor networks. By minimizing average AoI, systems can ensure that the information being used for decision-making is as fresh as possible, leading to more accurate and timely responses.

1.2.3 Peak Age of Information

The peak age of information (PAoI) was introduced in [8] as the value of the AoI achieved immediately before receiving an update, i.e., $A_j = Y_j + T_j$ in Fig. 1.2, which provides insights into the worst-case scenario for the freshness of information in a communication system. Note that the PAoI is a discrete stochastic process and the time average PAoI over the interval of observation $(0, \tau)$ is given as

$$\Omega_\tau = \frac{1}{N(\tau)} \sum_{i=1}^{N(\tau)} A_j, \tag{1.6}$$

where $N(\tau)$ is the number of samples which completed service by τ. Under the stationary ergodic processes of Y_j and T_j, the time average PAoI is given as

$$\Omega = \mathbb{E}[Y_j] + \mathbb{E}[T_j]. \tag{1.7}$$

In (1.7), the average PAoI avoids the computation of $\mathbb{E}[Y_jT_j]$ in the average of AoI, which significantly reduces the complexity of the AoI performance analysis.

The PAoI is particularly relevant in safety-critical systems or applications where delays in information updates can have severe consequences. For example, in autonomous vehicles systems, ensuring that the maximum AoI remains within acceptable limits is crucial to prevent accidents.

1.2.4 AoI Penalty

Apart from the AoI and the PAoI, the AoI penalty was introduced as a metric to measure the dissatisfaction level of outdated data [9, 10]. A general non-decreasing function, $v(\cdot)$, named the age penalty function, is utilized to characterize the importance level of packet freshness. Then the AoI penalty is defined as a random process, given by

$$\Xi(t) = v(\Delta(t)) = v(t - V(t)). \tag{1.8}$$

It is noted that keeping the average AoI penalty of the system small is equivalent to maintaining the high value of information. Given an AoI penalty process $\Xi(t)$, its time average is calculated using a sample average, similar to the calculation of average AoI, given as

$$\Xi_\tau = \frac{1}{\tau}\int_0^\tau \Xi(t)\mathrm{d}t \tag{1.9}$$

within the interval of observation $(0,\tau)$. Note that the average AoI is a special case of average AoI penalty, i.e., $\Xi(t) = \Delta(t)$, if $v(t) = t$. Since AoI penalty quantifies the cost or negative consequence associated with an increase in the AoI within the system, it is relevant in applications where the impact of outdated information on decision-making processes needs to be quantified and the form of AoI penalty function should be defined to characterize the impact. Moreover, if the AoI penalty function is an indicator function, i.e.,

$$v(t) = \mathbb{1}_{\{t>d\}} = \begin{cases} 1, & \text{if } t > d, \\ 0, & \text{otherwise,} \end{cases} \tag{1.10}$$

then the average AoI penalty is the AoI violation probability [11], which represents the fraction of time that the AoI exceeds a given threshold d. This metric provides insights into the reliability and consistency of information delivery in the system, which is valuable in applications where maintaining a certain level of freshness in information is required but occasional delays or inconsistencies are tolerable. For instance, in communication networks with limited bandwidth, it may be acceptable for the AoI to occasionally exceed a certain threshold as long as it remains within acceptable bounds for the majority of the time.

1.3 Literature Review

Since being introduced in [7], the concept of AoI has reaped a wide range of attention and interest. The management of AoI holds significant importance in a wide range of real-time applications. In these applications, effective management of AoI is essential for optimizing system performance, enhancing decision-making processes, and ensuring the reliability. For example, in monitoring applications, such as security monitoring and environmental sensing, maintaining up-to-date information is critical for timely detection and response to events. Effective AoI management ensures that decision-makers have access to the most current data, enabling proactive intervention and risk mitigation. In autonomous systems, maintaining current situational awareness is essential for safe and efficient operation. Minimizing AoI enables timely response to changing environmental conditions, detection of obstacles or hazards, and effective navigation in dynamic environments.

In this section, we explore the existing research development to provide a comprehensive and systematic review of the research conducted on the AoI. We further discuss the limitations of existing research and identify the research directions that require additional exploration and improvement.

1.3.1 AoI in Queue Theoretic Frameworks

The early research to employ the AoI as the performance metric for a state update delivery system was conducted through a simple queuing model. As an efficient mathematical methodology, the queuing theory has been adopted to analyze the AoI performance of systems with probabilistic events. The main system features and characteristics that need to be determined are (i) the number of sources, (ii) the packet arrival process, (iii) the service process, (iv) the queue discipline, (v) the number of servers, and (vi) the packet management strategy. Some early studies assessing such features and characteristics are summarized as follows:

In [12], the authors studied the AoI performance in a first-come-first-served (FCFS) system where three different queue models were considered, namely, $M/M/1$, $M/D/1$, and $D/M/1$. The main result of this seminal work is to investigate the optimal packet arrival rate, according to the serving rate, with respect to (w.r.t.) the AoI. Specifically, the authors pointed out that keeping the server in the idle state within 47% of the time is optimal w.r.t. the average AoI in the $M/M/1$ system. Apart from the deterministic and the specific exponentially distributed packet generation process and service process, a generally distributed packet generation process and service process has also been investigated. For example, a $D/G/1$ queue model was studied in [13] and a $G/G/1$ queue model was studied in [14]. Considering different numbers of servers, the authors of [15] investigated the AoI performance of the system with two servers and an infinite number of servers, i.e., $M/M/2$ and $M/M/\infty$ queue models. In addition, the trade-off between the AoI performance and the use of network resources has been demonstrated. For example, the analysis of the single-user system in [12] was extended to multi-user systems, where multiple users share a common queue [16, 17, 18]. The analysis illustrated the impact of the competition among different users for service resources on the AoI performance.

Instead of the FCFS queuing discipline, a last-come-first-served (LCFS) queuing discipline was proposed in [19], which was shown to achieve a lower average AoI than the FCFS queuing discipline. Moreover, [19] discussed two different packet management strategies for the LCFS system, namely, the preemption strategy and the non-preemption strategy. In the preemption strategy, when a new packet is generated at the source, it is allowed to replace the current packet in service. In contrast, for the non-preemption strategy, the new generated packet has to wait for the current packet in service to finish. Furthermore, [20] considered the LCFS with non-preemption strategy and investigated the effect of various buffer sizes on both the average AoI and the PAoI. Due to the analytical difficulty of the average AoI, [21] introduced a stochastic hybrid system (SHS) approach for AoI analysis. Aided with the proposed SHS approach, the average AoI of the LCFS queuing discipline in tandem queue models was investigated in [22, 23].

1.3.2 AoI Based Scheduling Policy Design

In multi-user systems, the scheduling policy plays a crucial role in determining the order in which packets are selected for transmission to optimize the AoI performance. The preliminary research work on AoI based scheduling policy design investigated and compared the AoI performance of age-independent scheduling policies. In [24], the average AoI of three different age-independent scheduling policies, i.e., round robin (RR), work-conserving non-collision, and random access (RA), was analyzed. The authors further claimed that the RR policy is the optimal packet arrival-independent scheduling policy to achieve the minimum average AoI. However, the packet arrival process of machine-type communications (MTC) may be very sporadic in many practical cases, such as the external event-trigger MTC [25]. This sporadic traffic nature makes the RR policy less attractive for achieving a better AoI performance. Indeed, regardless of the activity of users, reserving dedicated resources for all users leads to a long waiting time for some active users to transmit, which in turn increases the AoI. Against this background, the impact of sporadic traffic nature of MTC systems on packet arrival-independent scheduling policy design is required to be explored.

Apart from the AoI performance analysis of the age-independent scheduling policy, packet scheduling policies have been designed based on the AoI of each user to optimize the AoI performance of the system, starting with low-complexity scheduling policies design. The greedy algorithm based scheduling policy was designed to optimize the average AoI in [26, 27]. With formulating age minimization problem as a restless multi-armed bandit problem, a Whittle index approach was introduced in [28] for the scheduling policy design to minimize the average AoI in a multi-user system with a deterministic packet generation pattern and reliable transmission links. Similar to the greedy algorithm based scheduling policy, the Whittle index based scheduling policy gives each user an index and selects the user with largest index for scheduling. In contrast to the myopic and current impact focus of the greedy algorithm, the Whittle index approach is forward-looking, which considers not only the present impact on AoI performance but also the future impact of decisions. Inspired by [28], considering the practical signal propagation and packet generation pattern, [29, 30, 31] further extended the system model to a stochastic packet generation model with unreliable transmission links, and designed Whittle index based scheduling policies to minimize the average AoI. Motivated by the power of learning approaches for problem solving, the reinforcement learning and deep reinforcement learning based scheduling policies were designed in [32, 33, 34, 35, 36] to optimize the AoI performance. Although these designed scheduling policies almost achieve the optimal AoI performance, they require a significant amount of data to build accurate models and make informed decisions, since learning approaches make decisions based on past observations. This reliance on historical data can be problematic in dynamic and rapidly changing environments where characteristics and traffic patterns of systems may evolve quickly.

Building upon the efforts on the linear AoI performance, increasing research efforts have been devoted to designing the scheduling policy to optimize the general AoI penalty performance. In [37], a threshold based scheduling policy was proposed to optimize a general average AoI penalty in network control systems. Considering the deterministic packet generation model, [38] designed the Whittle index based scheduling policy to minimize the average AoI penalty. However, only the deterministic packet generation model was considered in [38], while the practice stochastic packet generation pattern remains unexplored. Against this background, it becomes essential to explore a low-complexity scheduling policy that minimizes the general average AoI penalty with stochastic packet generation and unreliable transmission links.

1.3.3 AoI in Mobile Edge Computing Systems

When data processing is computationally intensive, the user equipment (UE) with a limited computing capacity results in time-consuming data processing, which seriously degrades the AoI performance. To tackle this problem, mobile edge computing (MEC) was introduced by the European Telecommunications Standard Institute in 2015 [39]. In MEC systems, the server deployed at the network edge, called the edge server, is exploited to partially offload data processing tasks from the UE. As the MEC server usually has sufficient computing capacity, offloading can greatly reduce the computation time and save the UE's energy [40, 41].

For computation offloading, the main challenge is to determine whether or not to offload and how many data processing tasks should be offloaded [42]. The computation offloading decision is influenced by a number of parameters, such as the capacity of the MEC server, the number of UEs, as well as the efficiency of wireless packet transmission [43]. In the literature, significant attention has been given to the design of computation offloading strategies in MEC systems, aiming to minimize execution latency [44, 45, 46, 47]. Alongside reducing execution latency, other works have focused on additional objectives. For instance, [48, 49] aimed to maximize the energy efficiency, while [50] sought to optimize system scalability with the constraint of execution latency.

The AoI performance was firstly analyzed in point-to-point MEC systems. In [51], the AoI performance was analyzed in two data processing scenarios, i.e., the complicated characteristic extraction and classification in computer vision, and the optimal sampling and updating rate was found to minimize the average AoI. Considering single-user scenarios, the offloading policies were investigated in [52, 53]. Particularly, the authors of [53] proposed a partial offloading policy, where a proportion of computing tasks are offloaded to the edge server and the remaining tasks are processed by the local server. By carefully partitioning tasks allocation, the introduced partial offloading policy achieves the better AoI performance than other benchmark policies. Moreover, the impact of packet management strategies on the AoI performance was studied in [54, 55]. In particular, [54] studied the AoI performance of an MEC system with the preemption policy and the non-preemption policy. The authors of [55] designed a cutoff policy for an MEC system, where a fresh packet replaces the previous packet once the computation time of the previous packet is larger than a threshold. Apart from offloading policies and packet management strategies, [56] investigated the impact of different coding schemes on the AoI performance and found that the multiple computations coding scheme minimizes the average AoI of the MEC system.

Building upon the efforts on the MEC system with a single user, increasing research efforts have been devoted to investigating the AoI performance of multi-user MEC systems. For example, [57] derived the closed-form expression for the average PAoI and designed a min–max optimization problem to minimize the maximum average PAoI of an MEC system. The authors of [58] jointly designed a scheduling and status sampling policy to minimize the average AoI of the MEC system. In addition, [59] designed the edge resource allocation to minimize the average AoI of a multi-user MEC system.

Although the aforementioned studies have analyzed the AoI performance and designed the scheduling policy to optimize AoI performance in different MEC systems, the solution to the key problem: "How many computational tasks need to be offloaded in a multi-user MEC system such that its AoI performance is optimized?" has not been clearly presented. This necessitates further exploration and investigation.

1.3.4 AoI in Short Packet Communications

Due to the requirement of low latency in real-time applications, an unprecedented restriction on packet size is imposed. Indeed, short packet communications have been recognized as the potential solution for meeting the low latency requirement [60]. However, compared with the traditional long-packet communications, the transmission error in short packet communications is not negligible and highly depends on the blocklength of transmitted packets [61, 62].

Motivated by the benefits of short packet communications on latency reduction, the AoI performance of short packet communications was analyzed to evaluate the impact of short packets on the freshness of transmitted information. Specifically, [63] investigated the impact of the packet blocklength on the delay and the PAoI in a single-user system. Considering the same system, [64, 65] extended [63] to analyze the probability of the peak-age violation. Focusing on a decode-and-forward relaying system, [66] estimated the impact of the packet generation rate, the packet blocklength, and the blocklength allocation factor on the average AoI. Moreover, [67] studied the optimal packet blocklength of non-preemption and preemption strategies for minimizing the average AoI, while [68] derived the average AoI and the AoI violation probability in a downlink system. In [69], the impact of the packet blocklength on the average AoI of a multi-user system under two transmission scheduling policies, i.e., RR policy and maximum-age policy, was investigated. Considering a feedback mechanism, [70] derived the average AoI under two protocols, i.e., the traditional protocol and the automatic repeat request (ARQ) protocol, and determined the blocklengths to minimize the average AoI for both protocols. Furthermore, the authors of [71] considered a hybrid ARQ (HARQ) protocol scenario and investigated the optimal packet blocklength to minimize the average AoI. Focusing on the same system, a waiting policy was introduced in [72] to further improve the AoI performance.

Due to the strong interdependence between transmission reliability and latency on the packet blocklength, a critical problem arises concerning how to efficiently encode the collected information for transmission. Finding an effective encoding strategy becomes essential to improve AoI performance in short packet communication systems. In addition, it is essential to explore the impact of the packet blocklength on the AoI penalty performance to gain deep insights into the joint requirements of system reliability and timeliness.

1.3.5 AoI in Unmanned Aerial Vehicle Networks

Unmanned aerial vehicles (UAVs) are anticipated to play a vital role in future wireless networks, offering the potential to enable wireless communication and support high throughput [73]. Their integration into communication systems holds promise for enhancing connectivity and expanding the capabilities of wireless communication. A UAV-enabled network refers to a communication infrastructure that leverages UAVs as flying relays to facilitate long-range and extended coverage wireless connectivity. This architecture extends the reach of wireless networks, especially in scenarios where traditional terrestrial infrastructure is limited or unavailable [74]. In particular, UAVs act as mobile nodes that establish communication links between distant locations or areas with poor network connectivity, which serve as relay nodes that receive and transmit data packets between the source and destination, effectively creating a communication path. Comparing with satellite-enabled networks, UAV-enabled networks are more flexible and achieve lower propagation delay, which greatly improve the AoI performance.

In UAV-enabled networks, research interests primarily revolve two key aspects, i.e., resource allocation design and trajectory planning, to improve the the AoI performance. On one hand, resource allocation design involves investigating transmission budget allocation, including the transmission channel and power assignment for UEs and UAVs scheduling. To optimize the AoI performance, the scheduling policy was designed in [75, 76, 77]. In addition, leveraging the spectral efficiency and throughput improvement of the non-orthogonal multiple access (NOMA) technology [78], the transmission power allocation among multiple UEs was designed to optimize the AoI performance in [79, 80, 81]. On the other hand, the trajectory planning for UAVs involves determining the optimal path and motion of UAVs while considering various constraints and factors. The trajectory planning strategy was designed in [82, 83] to optimize the AoI performance, where the UAV transmits packets to the scheduled UE exclusively when it is directly positioned above the scheduled UE. Since the line-of-sight (LoS) propagation may increase the received signal power, [84] proposed a trajectory planning strategy to optimize the AoI performance with considering the impact of the altitude on the LoS probability. Considering the energy constraints of UAV and IoT UEs, the trajectory planning was investigated in [85, 86, 87, 88, 89, 90] to optimize the AoI performance. Given the complexity of trajectory planning design, which often leads to a non-deterministic polynomial-time (NP) problem, the deep reinforcement learning approach was employed in [91, 92, 93] to find the near optimal trajectory planning strategy.

The aforementioned studies have investigated the resource allocation design and trajectory planning in UAV-enabled networks with fixed networks where positions of UEs are fixed. However, the mobility of UEs may lead to dynamic networks in practice, where the design of resource allocation and trajectory planning in fixed networks cannot be applied into dynamic networks. This necessitates further exploration and investigation.

1.3.6 AoI Based Prototypes

Recently, several research works concentrated on the AoI based prototype design to test the AoI performance of medium access control (MAC) layer protocols on software-defined radio platforms (USRPs). These efforts aim to explore and validate the practical implementation of

AoI based protocols in real-world communication scenarios.

The authors of [94] developed and implemented a working testbed based on Ettus USRP devices, ensuring time synchronization among all devices by employing a slotted access scheme to access the channel medium. In [94], the designed prototype was utilized to test and compare the AoI performance of recently proposed AoI oriented random access policies. The experimental results demonstrated that the age-dependent random access policy performs better than the age-independent random access policy by achieving the lower average AoI, which corroborates well with existing analytical results. Inspired by [94], [95] implemented the prototype to evaluate four packet management and scheduling policies, i.e., the RA policy, the FCFS with tail drop policy, the FCFS with head drop policy, and the LCFS with packet discard policy. The experimental results showed that the the LCFS with packet discard policy outperforms the other three policies by achieving the lower average AoI. In addition, a WiFresh network architecture was proposed in [96], which combined the polling multiple access mechanism, the LCFS queuing policy, and the max-weighted scheduling policy. Prototypes of the proposed WiFresh network architecture, as well as the WiFi network architecture based on IEEE 802.11 standard [97], were implemented by using USRP devices. The experimental results revealed that the proposed WiFresh network architecture can significantly enhance the AoI performance compared to an equivalent standard WiFi network, particularly under high load conditions.

1.4 Thesis Outline and Contributions

This thesis primarily focuses on the AoI analysis and optimization in wireless networks. Specifically, we investigate the impact of different MAC layer and physical layer protocols, such as scheduling policies, packet management strategies, and encoding schemes on the AoI performance in various wireless networks, aiming to address the research gaps discussed in Section. 1.3.2, 1.3.3, and 1.3.4. We also explore the impact of the information correlation for UEs on the AoI based transmission protocol design in practical highly cooperative real-time systems. The specific contributions of each chapter are detailed below:

Chapter 2 – Scheduling Policy Design for AoI Optimization

In Chapter 2, we explore the scheduling policy design to improve the AoI performance in multi-user systems. In this chapter, we propose two different scheduling policies, i.e., a decentralized uplink packet arrival-independent scheduling policy, and a centralized downlink packet arrival-dependent scheduling policy. We present the main contributions of the two scheduling policies design separately in this chapter, where the main contributions of the decentralized uplink packet arrival-independent scheduling policy design are summarized as follows:

- We propose a novel scheduling policy to decrease the average AoI in an uplink multi-user MTC system. In particular, we design a repeated RA based scheduling policy, where UEs transmit packets to the BS via a shared transmission block (TB) following the RA policy, with a fixed number of repeated transmissions. Aided by simulation results, we show that our proposed scheduling policy outperforms the RR policy by achieving a lower average AoI, which demonstrates the benefits brought by the designed scheduling policy into MTC systems.

- We derive closed-form expressions for the average AoI of three different package management strategies, i.e., the FCFS with one buffer size, the LCFS with preemption in buffer, and the FCFS with infinite buffer size, in the proposed scheduling policy. Aided by simulation results, we validate our theoretical analysis and examine the impact of system parameters, such as the packet generation rate, the numbers of repeated transmissions and orthogonal frequency channels, and the transmission error probability, on the average AoI.

The main contributions of the centralized downlink packet arrival-dependent scheduling policy design are summarized as follows:

- We design a centralized packet arrival-dependent to minimize the general average AoI penalty in a downlink multi-user system with stochastic packet arrivals and unreliable channels. In particular, we first formulate the transmission scheduling problem as an average cost constrained Markov decision process (CMDP) problem and derive the approximated Whittle index. Based on the approximated Whittle index, we propose a transmission scheduling policy to minimize the general average AoI penalty.

- Our examination shows that our proposed policy achieves substantial AoI performance improvement compared to the existing policies, for given non-decreasing functions. We also investigate the impact of a heterogeneous scenario, where UEs have different dissatisfaction levels of the outdated data, on scheduling policy design and show that our proposed policy achieves a larger AoI performance gain in the heterogeneous scenario than the homogeneous scenario.

The results in this chapter have been presented in [98, 99], which is listed again for ease of reference:

[98] **Z. Tang**, Z. Sun, N. Yang, and X. Zhou, "Age of information of multi-source systems with packet management," in *Proc. IEEE Int. Conf. Commun. Workshop*, Dublin, Ireland, Jun. 2020, pp. 1-6.

[99] **Z. Tang**, Z. Sun, N. Yang, and X. Zhou, "Whittle index-based scheduling policy for minimizing the cost of age of information," *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 54-58, Jan. 2022.

Chapter 3 – AoI Analysis and Optimization of MEC Systems

In Chapter 3, we investigate how many computational tasks in each packet should be offloaded to the edge server, to optimize the AoI performance in a multi-user downlink MEC system. The main contributions of this chapter are summarized as follows:

- We derive closed-form expressions for the average AoI and the average PAoI of three computing schemes. Using simulations, we demonstrate the accuracy of our analysis results. We also find that the edge computing scheme achieves the lowest average AoI compared with the local computing scheme and the partial computing scheme when the packet generation rate and the number of UEs are small. We further find that by carefully partitioning the computational tasks, the partial computing scheme has a better system stability than the other two computing schemes.

- We derive the upper bound and the lower bound on the average AoI. With a sporadic packet generation pattern, we find that the average AoI is close to the average PAoI. We further obtain the optimal offload ratio to minimize the average PAoI. Simulations show that the gap between the optimal average AoI and the average AoI with the offloading ratio to minimize the average PAoI is negligible regardless of the packet generation rate. This optimal offloading ratio intuitively indicates what proportion of tasks should be offloaded to minimize the average AoI and PAoI in a multi-user MEC system.

- We investigate the impact of the system parameters on the optimal offloading ratio of the MEC system. We observe that the optimal offloading ratio is proportional to the computation rate of the edge server and inversely proportional to the computation rate of the local server and the number of UEs. We further find that the transmission rate hardly affects the optimal offloading ratio in the partial computing scheme.

The results in this chapter have been presented in [100, 101], which is listed again for ease of reference:
[100] **Z. Tang**, Z. Sun, N. Yang, and X. Zhou, "Age of information of multi-user mobile edge computing systems," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1600-1614, Jul. 2023.
[101] **Z. Tang**, Z. Sun, N. Yang, and X. Zhou, "Age of information analysis of multi-user mobile edge computing systems," in *Proc. IEEE Global Commun. Conf.*, Madrid, Spain, Dec. 2021, pp. 1-6.

Chapter 4 – AoI Analysis of Short Packet Communications

In Chapter 4, we investigate the AoI performance of short packet communication systems from two aspects. In particular, we first investigate the impact of different packet encoding schemes on the average AoI performance and then analyze the non-linear AoI penalty performance of short packet communication systems. We present the main contributions of the investigation on encoding schemes and non-linear AoI performance separately in this chapter, where the main contributions of the investigation on encoding schemes are summarized as follows:

- We derive closed-form expressions for the average AoI achieved by both encoding schemes, i.e., the joint encoding scheme and the distributed encoding scheme. Through these expressions, we identify the optimal blocklength of a packet that minimizes the average AoI for both encoding schemes. Aided by simulations, we demonstrate the accuracy of our analytical results. Moreover, we find that the block error rate has a more significant impact on the average AoI for the distributed encoding scheme compared to the joint encoding scheme. This insight highlights the importance of considering characteristics of the encoding scheme to optimize the AoI performance in short packet communication systems.
- We theoretically determine the condition when one encoding scheme has a better AoI performance than the other. Using simulations, we demonstrate the accuracy of our analytical results and discuss the appropriate option of the encoding scheme to decrease the average AoI based on information redundancy and the number of sensors. Simulation results also show that the distributed encoding scheme is more appropriate for systems with a relatively large number of sensors, compared with the joint encoding scheme.

The main contributions of the analysis on non-linear AoI performance are summarized as follows:

- We adopt a generalized α-β AoI penalty function to characterize the receiver's dissatisfaction level on outdated data and derive closed-form expressions for the average AoI penalty achieved by three packet management strategies, namely, the non-preemption with no buffer (NPNB) strategy, the non-preemption with one buffer (NPOB) strategy, and the preemption strategy, in a point-to-point short packet communication system. We also find the conditions of α to guarantee the existence of a finite average AoI penalty.
- We derive the closed-form expression for the average AoI penalty achieved by the zero-waiting policy and identify it as the lower bound on the average AoI penalty achieved by NPNB and NPOB strategies. We also evaluate the impact of various system parameters, such as the coding rate and status update generation rate, on the AoI penalty performance. Additionally, we discover that the value of α reflects the requirement of the system transmission reliability, providing valuable insights into optimizing AoI performance in practical communication scenarios.

The results in this chapter have been presented in [102, 103], which are listed again for ease of reference:

[102] **Z. Tang**, N. Yang, P. Sadeghi, and X. Zhou, "The age of information of short-packet communications: Joint or distributed encoding?" in *Proc. IEEE Int. Conf. Commun.*, Seoul, Republic of Korea, May 2022, pp. 2175-2180.
[103] **Z. Tang**, N. Yang, X. Zhou, and J. Lee, "Age of information penalty of short-packet communications with packet management," in *Proc. IEEE Int. Conf. Commun.*, Rome, Italy, May 2023, pp. 1-6.

Chapter 5 – AoI Based Transmission Scheme Selection

In Chapter 5, we study the AoI performance of a multi-user short packet communication system and the impact of the correlation among the information for users on different transmission schemes selection to optimize the AoI performance. The main contributions of this chapter are summarized as follows:

- We derive new closed-form expressions for the average AoI of a multi-user system, where we consider two transmission schemes with three packet management strategies, i.e., the non-preemption strategy, the preemption in buffer strategy, and the preemption in serving strategy. Aided by simulations, we demonstrate the accuracy of our analytical results. We also find that the block error rate has a more significant impact on the average AoI for the unicast transmission scheme than the broadcast transmission scheme. We further find that the non-preemption strategy achieves a lower average AoI than the preemption in serving strategy in the broadcast transmission scheme, and the preemption in buffer strategy achieves the lowest average AoI comparing with the non-preemption and the preemption in serving strategies in the unicast transmission scheme.

- Considering the information correlation among UEs, we compare the average AoI achieved by both transmission schemes with non-preemption strategy in a remote control system. In this system, with the feature of stochastic status update generation, we derive the threshold of the information ratio between the broadcast transmission scheme and the unicast transmission scheme to decide which transmission scheme is adopted. We then show the relationship between this threshold and the number of UEs in two special cases. We further find that the unicast transmission scheme achieves a lower average AoI than the broadcast transmission scheme when the number of UEs is large.

- We then extend the comparative study of these two transmission schemes to a dynamic system. In this system, we approximate the expected average AoI under the zero-waiting policy. Based on the approximation result, we determine the approximated threshold of the ratio between the individual information and the common information to decide which transmission scheme is adopted. We observe that the increase in this ratio has a more pronounced impact on the average AoI for the broadcast transmission scheme than the unicast transmission scheme.

The results in this chapter have been presented in [104, 105], which are listed again for ease of reference:

[104] **Z. Tang**, N. Yang, P. Sadeghi, and X. Zhou, "Age of information in downlink systems: Broadcast or distributed transmission?" *IEEE J. Select. Areas Commun.*, vol. 41, no. 7, pp. 2057-2070, Jul. 2023.

[105] **Z. Tang**, N. Yang, P. Sadeghi, and X. Zhou, "Broadcast versus distributed short-packet transmission: An age of information perspective," in *Proc. IEEE Int. Conf. Commun.*, Rome, Italy, May 2023, pp. 1-6.

Chapter 2

Scheduling Policy Design for AoI Optimization

2.1 Introduction

In this chapter, we propose two scheduling policies to improve the AoI performance in multi-user systems. Considering the sporadic traffic nature of MTC, we propose a repeated RA based scheduling policy to improve the AoI performance in Section 2.2. In Section 2.3, to explore a low-complexity scheduling policy that minimizes the general average AoI penalty with stochastic packet generation and unreliable transmission links, we adopt Whittle's approach and design a Whittle index based scheduling policy.

2.2 Repeated Random Access Based Scheduling Policy

In this section, we propose a decentralized uplink scheduling policy in a multi-user system to improve the average AoI performance. In our proposed policy, each UE generates its packet through an independent and identical Bernoulli process at the beginning of each time slot, and then randomly selects one channel from multiple orthogonal frequency channels to repeatedly transmit its packet across multiple time slots. We then analyze the average AoI performance of the system. Specifically, we derive new closed-form expressions for the average AoI achieved by the proposed scheduling policy for three packet management strategies. The simulation results corroborate our analysis and demonstrate that the proposed scheduling policy can achieve a lower average AoI than the existing policy for the considered packet management strategies. In addition, the impact of system parameters, such as the packet generation probability, the numbers of time slots and orthogonal frequency channels, and the channel error probability, on the average AoI is examined.

This section is organized as follows. In Subsection 2.2.1, the system model and the average AoI of the considered system are presented. The closed-form expressions for the average AoI of three queuing policies are derived in Subsection 2.2.2. The numerical results and summary

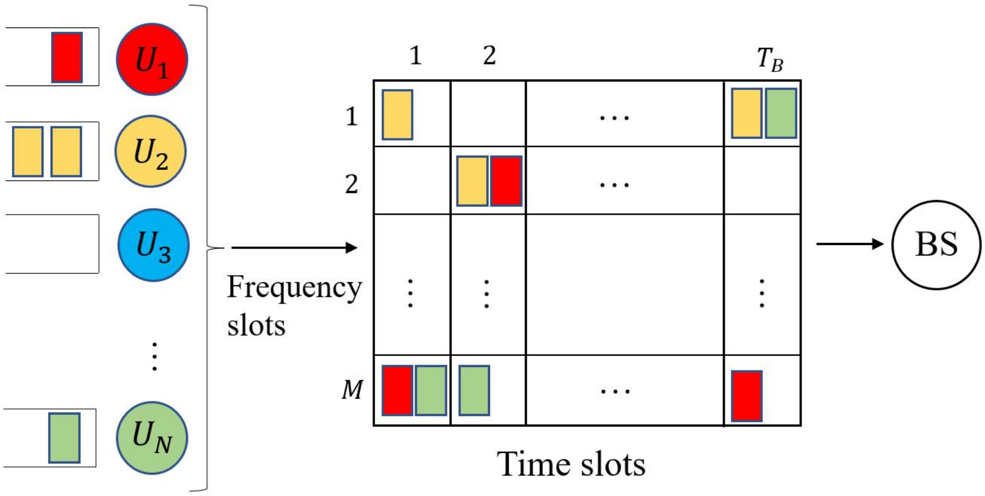


Figure 2.1: Illustration of our considered time-slotted system where N UEs transmit their status update packets to the BS via a shared TB.

are provided in Subsection 2.2.3 and 2.2.4, respectively.

2.2.1 System Model

We consider a time-slotted system as depicted in Fig. 2.1, where N UEs transmit their status update packets to a BS. In particular, N UEs, denoted by $\Phi_U = \{U_1, U_2, \cdots, U_N\}$, which have the homogeneous packet generation pattern transmit their generated packets to the BS via a shared TB. The TB consists of M orthogonal frequency channels and T_B time slots.

Each UE generates a packet at the beginning of each time slot with a probability λ independently and then stores it in the queue until the beginning of a new TB. When a UE attempts to transmit its packet at the beginning of a TB, this UE is referred to as an active UE and allowed to access the channel through an RA policy. In particular, at the beginning of each time slot in the TB, an active UE randomly, but following a uniform distribution, selects one channel from M orthogonal frequency channels for its transmission[1]. After this selection, we assume that this packet is repeatedly transmitted across all the T_B time slots[2]. A collision occurs if there are two or more packets transmitted in one channel simultaneously. Furthermore, we assume a packet error probability ε for investigating collision-free packet transmission, where the packet error is used to model the effect of practical channels on the transmitted short-length packets.

We assume that the BS successfully detects a packet only when the packet is not collided with other packets and does not experience erroneous transmission, and refer to this packet as the determinant packet. Otherwise, the transmission fails and the packet is dropped. If

[1]For example, in the slotted Aloha protocol, time is divided into discrete slots, and devices are only allowed to transmit at the beginning of each time slot.
[2]Note that the repetition transmission is considered in this section to eliminate the feedback overhead.

the BS successfully detects at least one replica of a UE's packet during the TB, it is believed that the packet successfully reaches the BS and the UE's status is updated. We denote $u_n(t)$, $n \in \{1, 2, \ldots, N\}$, as the generation time of the most recently detected packet from U_n at time slot t. The AoI of U_n at the time slot t is thus given by

$$\Delta_n(t) = t - u_n(t). \tag{2.1}$$

In this system, we consider that T time slots are grouped into K TBs and each TB has T_B time slots, i.e., $T = KT_B$. By averaging $\Delta_n(t)$ over all UEs and the entire time duration, we obtain the average AoI of the system as

$$\Delta = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{N} \sum_{n=1}^{N} \Delta_n(t) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{T_B} \sum_{\tau=1}^{T_B} \underbrace{\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \Delta_n\left((k-1)T_B + \tau\right)}_{\Delta_{n,\tau}}, \tag{2.2}$$

where $\tau = 1, 2, \cdots, T_B$. Without loss of generality, we arbitrarily select UE U_n as the *typical UE* and analyze its average AoI performance to evaluate the average AoI for the system. As all UEs share the same packet generation probability and transmission protocol, they have an identical AoI. Thus, the average AoI for the system in (2.2) can be written as

$$\Delta = \frac{1}{T_B} \sum_{\tau=1}^{T_B} \Delta_{n,\tau}. \tag{2.3}$$

In this work, we consider three packet management strategies in the queues with various buffer sizes. They are described as follows:

- FCFS-1: The queue of each UE stores only one packet and a newly generated packet is dropped if there is already one packet in the queue.
- LCFS-1: The queue of each UE stores only one packet but a newly generated packet replaces the previous packet in the queue. As a result, only the latest generated packet is stored in the queue.
- FCFS-∞: The queue of each UE has an infinite-size buffer. All of generated packets are stored in the queue and the service order of all the stored packets follows an FCFS manner.

2.2.2 Average AoI Analysis

In this section, we first analyze the average AoI for the proposed policy. Based on this result, we derive closed-form expressions for the average AoI by considering three packet management strategies.

We observe from (2.2) that the average AoI for the proposed policy can be obtained from the average AoI of a UE at each time slot. In order to derive the average AoI of a UE at each time slot during the TB, we first analyze the Δ_{n,T_B}. According to the system model, it is known that when a packet is generated by a UE, it needs to wait in the UE's queue until the beginning of a new TB. In particular, if the UE U_n transmits a packet in the kth TB, defined as the event $\mathcal{A}_{n,k}$, the waiting time of this packet is $W_{n,k} = (k-1)T_B - V_{n,k}$, where $V_{n,k}$ denotes the generation time of this packet. Otherwise, there is no packet in U_n's queue and the waiting time is zero. Thus, we express the waiting time of a packet as

$$W_{n,k} = \begin{cases} (k-1)T_B - V_{n,k}, & \mathcal{A}_{n,k}, \\ 0, & \text{otherwise}. \end{cases} \tag{2.4}$$

It is worthwhile to note that if U_n transmits a packet in the first TB, i.e., $k=1$, this packet is generated before the time of zero, and its waiting time is from the generation time to the time of zero, i.e., $V_{n,1} \le 0$ and $W_{n,1} = -V_{n,1}$.

When U_n transmits a packet and the BS successfully detects this packet in kth TB, i.e., the event $\mathcal{C}_{n,k}$ happens, the AoI of U_n at the end of this TB is $\Delta_n(kT_B) = W_{n,k} + T_B$. Otherwise, the AoI of U_n at the end of this TB is $\Delta_n(kT_B) = \Delta_n((k-1)T_B) + T_B$. Thus, the AoI of U_n at the end of this TB is given by

$$\Delta_n(kT_B) = \begin{cases} W_{n,k} + T_B, & \mathcal{C}_{n,k}, \\ \Delta_n((k-1)T_B) + T_B, & \text{otherwise}. \end{cases} \tag{2.5}$$

By exploiting the probability that the event $\mathcal{C}_{n,k}$ happens, denoted by $\Pr(\mathcal{C}_{n,k})$, we obtain

$$\mathbb{E}[\Delta_n(kT_B)] = T_B + \Pr(\mathcal{C}_{n,k})W_{n,k} + (1-\Pr(\mathcal{C}_{n,k}))\Delta_n((k-1)T_B). \tag{2.6}$$

The probability $\Pr(\mathcal{C}_{n,k})$ in (2.6) is derived as

$$\begin{aligned} \Pr(\mathcal{C}_{n,k}) &= \Pr(\mathcal{C}_{n,k},\mathcal{A}_{n,k}) = \Pr(\mathcal{A}_{n,k})\Pr(\mathcal{C}_{n,k}|\mathcal{A}_{n,k}) = \Pr(\mathcal{A}_{n,k})\sum_{n'=0}^{N-1}\Pr(\mathcal{B}_{n',k})\Pr(\mathcal{C}_{n,k}|\mathcal{A}_{n,k},\mathcal{B}_{n',k}) \\ &= P_a \sum_{n'=0}^{N-1}\binom{N-1}{n'} P_a^{n'}(1-P_a)^{N-1-n'}\Psi(n',T_B), \end{aligned} \tag{2.7}$$

where

$$\Psi(\mu,\nu) = 1 - \left(1 - \left(1-\frac{1}{M}\right)^{\mu}(1-\varepsilon)\right)^{\nu}. \tag{2.8}$$

In (2.7), $\mathcal{B}_{n',k}$ denotes the event that there are n' UEs, out of $N-1$ UEs, transmitting the packet in the kth TB and P_a denotes the update probability which evaluates the likelihood that the event $\mathcal{A}_{n,k}$ happens. Due to the independent and identical packet generation process for N UEs, P_a is same for all UEs.

When the system reaches the steady-state, we have

$$\Pr(\mathcal{C}_{n,k}) = \Pr(\mathcal{C}_{n,k+1}), \forall k \tag{2.9}$$

and

$$\begin{aligned}\Delta_{n,T_B} &= \mathbb{E}[\Delta_n(kT_B)] = \mathbb{E}[\Delta_n((k+1)T_B)] \\ &= T_B \left(P_a \sum_{n'=0}^{N-1} \binom{N-1}{n'} P_a^{n'} (1-P_a)^{N-1-n'} \Psi(k, T_B) \right)^{-1} + W_n,\end{aligned} \tag{2.10}$$

where the expectation is taken over all the K TBs. In (2.10), W_n is the average packet waiting time in the queue of U_n, which can be computed as

$$W_n = \lim_{K \to \infty} \frac{\sum_{k=1}^{K} W_{n,k}}{\sum_{k=1}^{K} \mathbb{1}(\mathcal{A}_{n,k})}. \tag{2.11}$$

Based on the analysis of Δ_{n,T_B}, we next derive $\Delta_{n,\tau}$, where $\tau = 1, 2, \cdots, T_B - 1$. At the BS, we denote $\mathcal{C}_{n,k}^{\tau}$ as the event that the BS successfully detects a packet of U_n during the first τ time slots of the kth TB. Similar to the derivation of $\Pr(\mathcal{C}_{n,k})$, the probability $\Pr(\mathcal{C}_{n,k}^{\tau})$ is computed as

$$\Pr(\mathcal{C}_{n,k}^{\tau}) = P_a \sum_{n'=0}^{N-1} \binom{N-1}{n'} P_a^{n'} (1-P_a)^{N-1-n'} \Psi(n', \tau). \tag{2.12}$$

Moreover, similarly to obtain (2.5), the AoI of U_n at the τth time slot in kth TB can be derived as

$$\Delta_n((k-1)T_B + \tau) = \begin{cases} W_{n,k} + \tau, & \mathcal{C}_{n,k}, \\ \Delta_n((k-1)T_B) + \tau, & \text{otherwise}. \end{cases} \tag{2.13}$$

Based on (2.12) and (2.13), we obtain

$$\mathbb{E}[\Delta_n((k-1)T_B + \tau)] = \tau + \Pr(\mathcal{C}_{n,k}^{\tau}) W_{n,k} + (1 - \Pr(\mathcal{C}_{n,k}^{\tau})) \Delta_n((k-1)T_B). \tag{2.14}$$

By averaging over all the K TBs, we obtain $\Delta_{n,\tau}$, where $\tau = 1, 2, \cdots, T_B - 1$, as

$$\Delta_{n,\tau} = (1 - P_{c,\tau}) \Delta_{n,T_B} + P_{c,\tau} W_n + \tau, \tag{2.15}$$

where Δ_{n,T_B} is derived in (2.10). Moreover, $P_{c,\tau}$ is the average probability that the event $\mathcal{C}_{n,k}^{\tau}$ happens, given by

$$P_{c,\tau} = \Pr\left(\mathcal{C}_{n,k}^{\tau}\right) = \Pr\left(\mathcal{C}_{n,k+1}^{\tau}\right), \tag{2.16}$$

which is due to the independent and identical packet generated process of N UEs. Based on the expression for the average AoI in (2.2), we obtain the desired AoI result, $\Delta_{n,\tau}$ with $\tau = 1, 2, \cdots, T_B$, by combining the expressions in (2.10) and (2.15).

We note that both the update probability of a UE and the average waiting time of a packet are associated with the employed packet management strategy and the buffer size. Thus, we next analyze P_a and W_n in three different packet management strategies in the following subsections.

2.2.2.1 FCFS-1 and LCFS-1 Systems

In this subsection, we first derive the update probability of a UE and the average waiting time of a packet in the FCFS-1 system. Then we derive them in the LCFS-1 system.

In the FCFS-1 system, only the packet generated during the kth TB is transmitted to the BS at the $(k+1)$th TB. Thus, the update probability P_a in the FCFS-1 system is given by

$$P_{a,F} = 1 - (1-\lambda)^{T_B}, \tag{2.17}$$

and the probability mass function (PMF) of the generation time of a packet within a TB is given by

$$\Pr\left(V_{n,k}^{F} = t | \mathcal{A}_{n,k}\right) = \frac{\lambda (1-\lambda)^{t-1}}{P_{a,F}}, \tag{2.18}$$

where $V_{n,k}^{F}$ is $V_{n,k}$ for the FCFS-1 system. Based on (2.18), (2.4), and (2.11), we obtain the average waiting time of a packet as

$$W_n^F = \frac{\lambda}{P_{a,F}} \sum_{t=1}^{T_B} (1-\lambda)^{t-1} (T_B - t) = \left(\frac{T_B - 1}{P_{a,F}} - \frac{(1-\lambda)\left(1-(1-\lambda)^{T_B-1}\right)}{\lambda P_{a,F}} \right), \tag{2.19}$$

where W_n^F is W_n for the FCFS-1 system. By integrating (2.17) and (2.19) into (2.10), we obtain Δ_{n,T_B} in the FCFS-1 system. Then, by integrating (2.10) into (2.15) and (2.2), the average AoI of the FCFS-1 system is obtained.

Similarly to the FCFS-1 system, there is only one packet stored in the queue for the LCFS-1 system. Thus, the update probability is computed as

$$P_{a,L} = 1 - (1-\lambda)^{T_B}, \tag{2.20}$$

which is the same as (2.17). Since the latest packet in the queue is served first in the LCFS-1 system, the PMF of the generation time of a packet within a TB is given by

$$\Pr\left(V_{n,k}^{L}=t|\mathcal{A}_{n,k}\right)=\frac{\lambda(1-\lambda)^{T_B-t}}{P_{a,L}}, \tag{2.21}$$

where $V_{n,k}^{L}$ is $V_{i,k}$ for the LCFS-1 system. Then, the average waiting time of a packet is obtained as

$$W_n^L=\frac{\lambda}{P_{a,L}}\sum_{t=1}^{T_B}(1-\lambda)^{T_B-t}(T_B-t)=\left(\frac{(1-\lambda)\left(1-(1-\lambda)^{T_B-1}\right)}{\lambda P_{a,L}}-\frac{(T_B-1)(1-\lambda)^{T_B}}{P_{a,L}}\right), \tag{2.22}$$

where W_n^L is W_n for the LCFS-1 system. By integrating (2.20) and (2.22) into (2.10), we obtain Δ_{n,T_B} in the LCFS-1 system. Then by integrating (2.10) into (2.15) and (2.2), the average AoI of the LCFS-1 system is obtained.

2.2.2.2 FCFS-∞ System

In this subsection, we consider the FCFS system with an infinite-size buffer, i.e., the FCFS-∞ system. For the FCFS-∞ system, since there are two or more packets in the queue, the update probability and the average waiting time are associated with the queue state, i.e., the number of packets in the queue. Then, we first analyze the queue state as a Markov process.

Due to the same pattern of N UEs, the queue state of each UE follows an independent and identical Markov process. Without loss of generality, we denote $\pi_{T_B^+}$ as the queue state of the typical UE, U_n, at the beginning of the TB after a new packet transmitted and $\pi_{T_B^+}(k)=s$ as the queue containing s packets at the beginning of kth TB. We note that only when the packet generating rate is less than its serving rate, the queue state is stable, i.e., $T_B\lambda<1$. We denote $\mathbf{p}_{T_B^+}$ as the stationary probability distribution of $\pi_{T_B^+}$ and $p_{T_B^+}(s)$ as the stationary probability of that there are s packets in the queue at beginning of the TB after a new packet being transmitted. Based on the state transition probability of the queue, we obtain $p_{T_B^+}(s+1)$ as

$$p_{T_B^+}(s+1)=\begin{cases}\frac{p_{T_B^+}(0)-P_1p_{T_B^+}(0)}{P_0}-P_1p_{T_B^+}(0), & s=0,\\ \frac{p_{T_B^+}(s)-\sum_{t=1}^{T_B}P_tp_{T_B^+}(s+1-t)}{P_0}, & s=1,2,\cdots,\end{cases} \tag{2.23}$$

where P_t is the probability that there are t packets generated in one TB, which is given by

$$P_t=\binom{T_B}{t}\lambda^t(1-\lambda)^{T_B-t}. \tag{2.24}$$

We denote $G_{T_B^+}(Z)$ as the probability generating function (PGF) of $p_{T_B^+}$, which is given by

$$G_{T_B^+}(Z)=\sum_{s=0}^{\infty}p_{T_B^+}(s)Z^s=p_{T_B^+}(0)(1-Z)+\frac{Z}{P_0}G_{T_B^+}(Z)-\frac{1}{P_0}\left(A(Z)-P_0\right)G_{T_B^+}(Z), \quad (2.25)$$

where $A(Z)$ is the PGF of the geometric distribution with the parameter λ, expressed as

$$A(Z)=(1-\lambda+\lambda Z)^{T_B}. \quad (2.26)$$

By integrating (2.26) into (2.25), we simplify the expression for $G_{T_B^+}(Z)$ as

$$G_{T_B^+}(Z)=\frac{P_0 p_{T_B^+}(0)(1-Z)}{(1-\lambda+\lambda Z)^{T_B}-Z}. \quad (2.27)$$

Based on the property of the PGF [106], we obtain

$$G_{T_B^+}(Z)|_{Z=\lim_{\delta\to 0}1-\delta}=1. \quad (2.28)$$

Then, the expression for $G_{T_B^+}(Z)$ can be further simplified as

$$G_{T_B^+}(Z)=\frac{(1-T_B\lambda)(1-Z)}{(1-\lambda+\lambda Z)^{T_B}-Z}. \quad (2.29)$$

We then analyze the queue state at the beginning of the TB before a new packet being transmitted and denote its stationary distribution as $p_{T_B^-}$. Specifically, the relationship between $p_{T_B^-}$ and $p_{T_B^+}$ is formulated as

$$p_{T_B^+}(s)=\begin{cases}p_{T_B^-}(0)+p_{T_B^-}(1), & s=0,\\ p_{T_B^-}(s+1), & s=1,2,\cdots,\end{cases} \quad (2.30)$$

and

$$p_{T_B^-}(0)=\frac{P_0}{1-P_0}p_{T_B^-}(1). \quad (2.31)$$

From (2.29), (2.30), and (2.31), we obtain $p_{T_B^-}(0)=1-T_B\lambda$. Thus, the update probability of each UE in the FCFS-∞ system is given by

$$P_{a,\infty}=1-p_{T_B^-}(0)=T_B\lambda. \quad (2.32)$$

We next derive the average waiting time in the queue of S_i in the FCFS-∞ system, denoted by W_i^{∞}. As there is an infinite buffer size to be considered, the waiting of a packet is caused by the transmission of previously generated packets. These packets include both the packets generated in previous TBs and the packets generated in the same TB. For the packets generated in previous TBs, they may lead to the waiting time of the packet as

$$W_m^{\infty,1} = \mathbb{E}[s_q T_B] = T_B \mathbb{E}[\pi_{T_B^+}] = \frac{T_B^2(T_B-1)\lambda^2}{2(1-T_B\lambda)}, \tag{2.33}$$

where s_q is the number of packets which are waiting to be served in the queue but generated during the previous TB when the determinant packet enqueues. For the packets generated in the same TB as the determinant packet, they may lead to the waiting time of the packet as

$$\begin{aligned} W_n^{\infty,2} = \mathbb{E}[s_g T_B + (T_B - \tau)] &= \sum_{\tau=1}^{T_B} \sum_{s_g=0}^{\tau-1} \Pr(s_g,\tau)(s_g T_B + (T_B-\tau)) \\ &= \frac{T_B-1}{2} + \frac{T_B(T_B-1)}{2}\lambda, \end{aligned} \tag{2.34}$$

where s_g is the number of packets which are generated before the determinant packet generated in the same TB and $\Pr(s_g,\tau)$ is the joint probability distribution of s_g and τ, given by

$$\Pr(s_g,\tau) = \frac{1}{T_B}\binom{\tau-1}{s_g}\lambda^{s_g}(1-\lambda)^{\tau-1-s_g}. \tag{2.35}$$

By jointly considering $W_n^{\infty,1}$ and $W_n^{\infty,2}$, the average waiting time of a packet for the FCFS-∞ system is obtained as

$$W_n^{\infty} = W_n^{\infty,1} + W_n^{\infty,2}. \tag{2.36}$$

By integrating (2.32) and (2.36) into (2.10), we obtain Δ_{n,T_B} in the FCFS-∞ system. Then by integrating (2.10) into (2.15) and (2.2), the average AoI of the FCFS-∞ system is obtained.

2.2.3 Numerical Result and Discussion

In this section, we present numerical results to validate our theoretical analysis in Section 2.2.2. Moreover, we evaluate the impact of various parameters on the average AoI of the proposed policy, including the packet generation probability, the number of time slots and the number of orthogonal frequency channels within a TB, and the channel error probability.

Fig. 2.2 plots the average AoI versus the packet generation probability λ. We first observe that the analytical result for the average AoI tightly matches Monte Carlo simulations. Then we observe that when λ increases, the average AoI first decreases and then increases. This implies that there is an optimal value of λ to minimize the average AoI. This observation is due to the fact that the increase in λ results in a two-fold effect on the average AoI via the update probability and the collision probability. In particular, when λ increases from a very low value, the increasing update probability dominantly and positively affects the average AoI. When the value of λ exceeds a certain threshold, the collision probability increases significantly, thereby

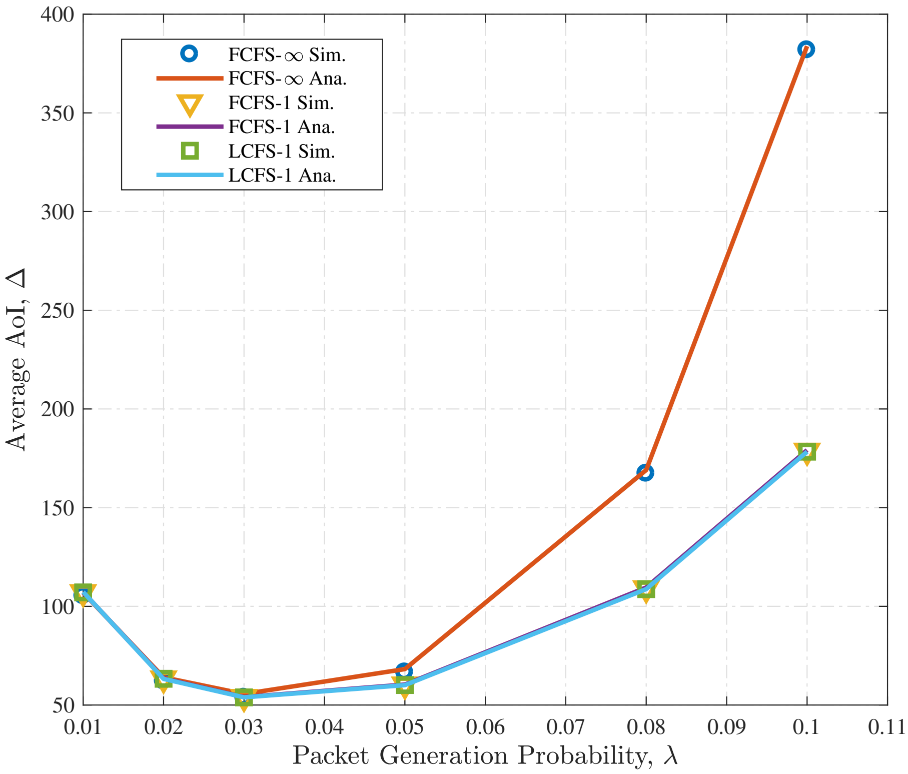


Figure 2.2: The average AoI versus the packet generation probability λ with $N = 40$, $M = 4$, $T_B = 5$, and $\varepsilon = 0.01$.

degrading the AoI performance. We further observe that the average AoI in FCFS-1 and LCFS-1 systems are almost the same and the advantage of the LCFS-1 system over the FCFS-1 system is not obvious. This is due to the fact that we have chosen a very small T_B. In addition, we observe that the gap between the average AoI in FCFS-∞ and FCFS-1 systems increases when λ increases. This is because both the update probability and the collision probability in the FCFS-∞ system increases faster than those in the FCFS-1 system.

Fig. 2.3 plots the average AoI versus the number of time slots within a TB, T_B. We observe from the figure that when T_B increases, the average AoI first decreases and then increases. This is due to the fact that on one hand, using more time slots in a TB leads to a higher successful probability for a UE, which in turn decreases the average AoI. On the other hand, using more time slots in a TB results in a long waiting time of a packet in the queue, which increases the average AoI. Then, we observe that the optimal number of time slots within a TB can be obtained to minimize the average AoI for FCFS-1 and LCFS-1 systems. Moreover, by comparing the average AoI of FCFS-1, LCFS-1, and RR-1 systems, we find that both the LCFS-1 system and the FCFS-1 system exhibit a better AoI performance than the RR-1 system

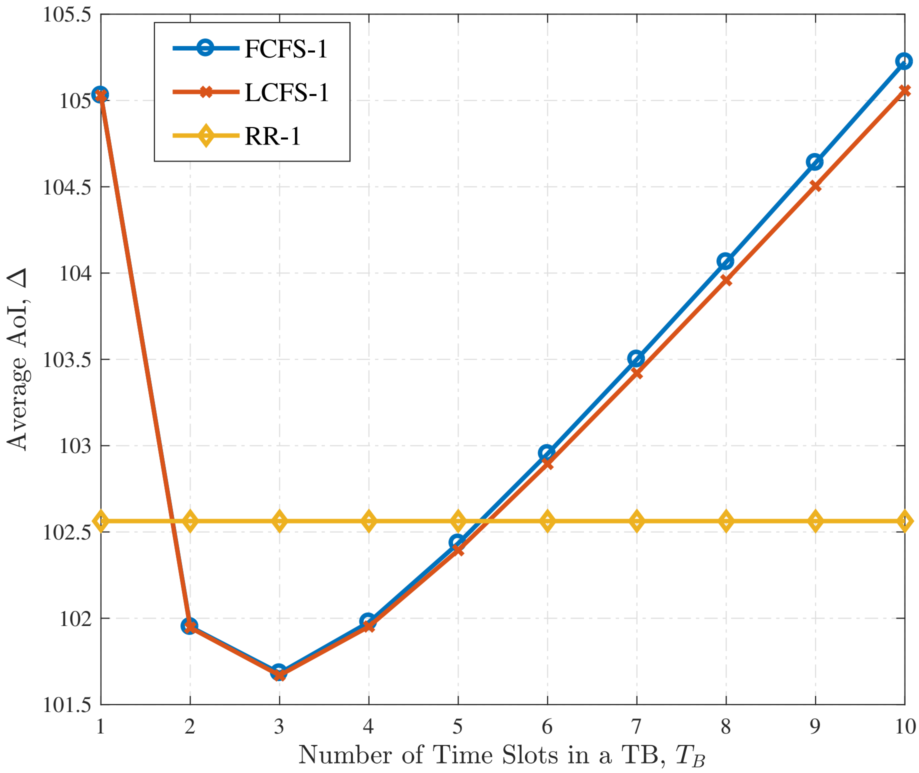


Figure 2.3: The average AoI versus the number of time slots in a TB, T_B, with $N = 40$, $\lambda = 0.01$, $M = 10$, and $\varepsilon = 0.01$.

under a particular range of T_B. This is because of that the RR policy is less efficient for sporadic traffic systems. Furthermore, we find that the average AoI gap between the FCFS-1 system and LCFS-1 system increases when T_B increases, which clearly shows the advantage of the LCFS-1 system.

Fig. 2.4 plots the average AoI of FCFS-1 and RR-1 systems versus the number of channels within a TB, M. We first observe that the average AoI of both FCFS-1 and RR-1 systems decreases when M increases. Moreover, we observe that when the channel error probability increases, the average AoI of the RR-1 system dramatically increases while the average AoI of the FCFS-1 system slightly increases. This demonstrates that in terms of the average AoI, the proposed system is more robust to channel errors than the RR-1 system.

2.2.4 Summary

In this section, we proposed an RA based scheduling policy which aims at improving the AoI performance of a multi-user system. In this system, the packet of each UE was generated by an independent and identical Bernoulli process at the beginning of each time slot and transmitted

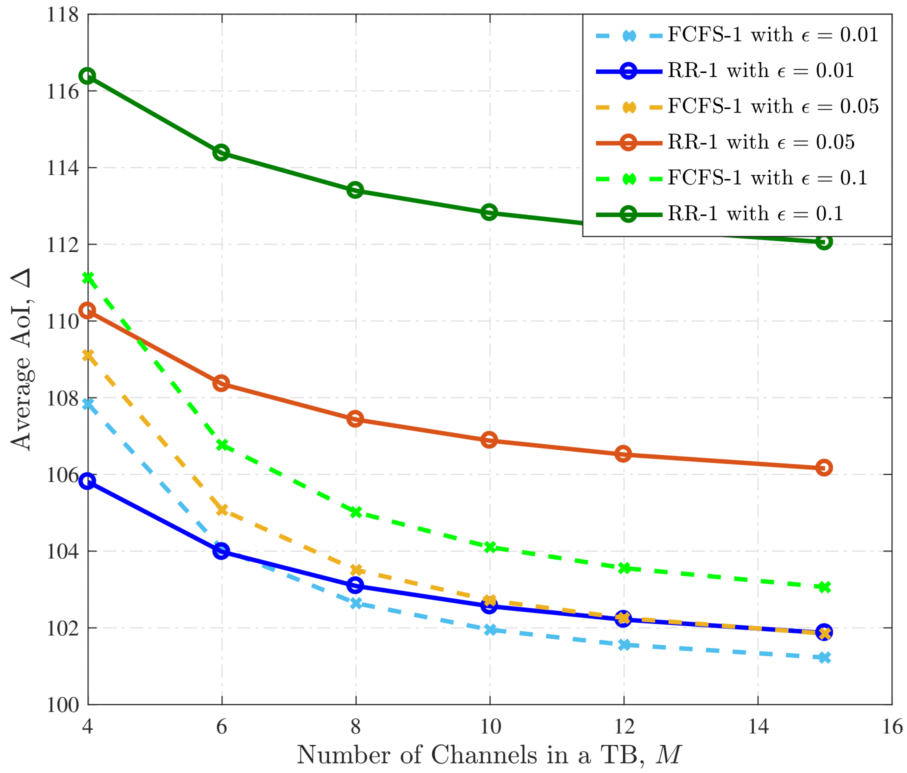


Figure 2.4: The average AoI versus the number of channels in a TB, M, with $\lambda = 0.01$, $T_B = 3$, and $N = 40$.

to the BS via a shared TB, which consists of M orthogonal frequency channels and T_B time slots. We derived closed-form expressions for the average AoI for three packet management strategies, i.e., FCFS-1, LCFS-1, and FCFS-∞. With numerical results, we demonstrated the accuracy of our analysis and examined the performance advantage of our proposed scheduling policy relative to the existing ones. Importantly, we found the optimal number of repeated transmission times to minimize the average AoI.

2.3 Whittle Index Based Scheduling Policy

In this section, we design a centralized downlink scheduling policy to minimize the general non-decreasing penalty function of the AoI in a multi-user system. In this system, the BS stochastically generates time-sensitive packets and transmits them to corresponding UEs via an unreliable channel. We first formulate the transmission scheduling problem as an average cost CMDP problem. By introducing a constant service charge, we derive the approximated expression for the Whittle index, based on which we design the scheduling policy. Using numerical results, we demonstrate the performance gain of our designed scheduling policy compared to the existing policies, such as the optimal policy, the on-demand Whittle index policy, and the age greedy policy.

This section is organized as follows. In Subsection 2.3.1, the system model and the average AoI penalty of the considered system are presented. The Whittle index based scheduling policy is designed in Subsection 2.3.2. The numerical results and summary are provided in Subsection 2.3.3 and 2.3.4, respectively.

2.3.1 System Model and Problem Formulation

We consider a time-slotted multi-user system, where the BS transmits packets to N different UEs, denoted by $\Phi_U = \{U_1, U_2, \cdots, U_N\}$. We assume that there are N buffers at the BS, each of which corresponds to one UE. At the beginning of each time slot, the BS generates the packet of U_n with a probability λ_n, $n \in \{1, 2, \ldots, N\}$[3]. When the packet of U_n is generated, it is stored in its corresponding buffer. We further assume that each buffer stores one packet such that the newly generated packet replaces the previous one in the buffer.

We assume that the BS schedules at most one packet transmission in each time slot, and hence, only the packet of one UE can be transmitted at a time. We denote a binary variable $\zeta_n(t) = \{0, 1\}$ as the scheduling indicator of U_n during time slot t. If U_n is scheduled for transmission during time slot t, $\zeta_n(t) = 1$. Otherwise, $\zeta_n(t) = 0$. Thus, we obtain

$$\sum_{n=1}^{N} \zeta_n(t) \leq 1, \forall t. \tag{2.37}$$

We assume that during each time slot, packet transmission occurs prior to new packet generation. Thus, only the packets generated in previous time slots can be transmitted by the BS.

By considering practical signal propagation between the BS and UEs, the transmitted packets may not be successfully detected by UEs. We denote a binary variable $\delta_n(t) = \{0, 1\}$ as an indicator of whether or not U_n successfully detects its packet transmitted by the BS during time slot t. If the packet is successfully detected by U_n during time slot t, $\delta_n(t) = 1$. Otherwise, $\delta_n(t) = 0$. The transmission error probability of U_n is denoted by ε_n, i.e., $\Pr(\delta_n(t) = 0) = \varepsilon_n$. When U_n successfully detects a packet, it immediately sends back an acknowledgment (ACK) to the BS via an error-free control channel. The feedback overhead is assumed to be negligible as the ACK length is much smaller than the packet length. Once an ACK is received from U_n, the BS empties the corresponding buffer by dropping the successfully received packet. Otherwise, this packet is held in the buffer.

We denote $V_n(t)$ as the generation time of the last successfully detected packet at U_n during

[3] We assume that the packet generation process is independent but not identical among UEs.

time slot t. Then, the AoI of U_n during time slot t is given by

$$h_n(t) = t - V_n(t). \tag{2.38}$$

Let us define $W_n(t)$ as the generation time of the newest generated packet of U_n. Until time slot t, the queuing delay of this packet is computed as $a_n(t) = t - W_n(t)$. The duration between the generation time of the last successfully detected packet and that of the newest generated packet is $d_n(t) = W_n(t) - V_n(t)$. Then, we express the AoI of U_n as $h_n(t) = a_n(t) + d_n(t)$. Thus, the evolution of the AoI of U_n is given by

$$h_n(t+1) = \begin{cases} a_n(t)+1, & \text{if } \zeta_n(t)\delta_n(t) = 1, \\ h_n(t)+1, & \text{otherwise.} \end{cases} \tag{2.39}$$

It is noted that when the BS successfully transmits a packet to U_n but does not generate a new packet, this successfully transmitted packet is considered as the newest generated packet. When this happens, we have $d_n(t) = 0$ and $h_n(t) = a_n(t)$.

Based on (2.38), we define the average AoI penalty as

$$\Xi = \lim_{T\to\infty} \frac{1}{TN} \sum_{t=1}^{T} \sum_{n=1}^{N} v(h_n(t)), \tag{2.40}$$

where $v(h_n(t))$ is a general non-decreasing function of AoI, characterizing the importance level of packet freshness. We observe from (2.39) that the AoI is determined by the adopted transmission scheduling policy, i.e., $\zeta_n(t)$. Thus, we can optimize the transmission scheduling policy to minimize the average AoI penalty in (2.40). The optimization problem is formulated as

$$\begin{aligned} \Xi_c = \min_{c\in\mathcal{C}} \lim_{T\to\infty} \frac{1}{NT} \sum_{t=1}^{T} \sum_{n=1}^{N} v(h_n(t)), \\ \text{s.t.} \quad \sum_{n=1}^{N} \zeta_n(t) \leq 1, \forall t, \end{aligned} \tag{2.41}$$

where $\mathcal{C}$ denotes the set of all potential transmission scheduling policies. In fact, this optimization problem is an infinite time horizon average cost CMDP problem, with the state space $\{(a_1(t), d_1(t)), \cdots, (a_N(t), d_N(t))\}$. We note that, due to the countably infinite state space of (2.41), it is infeasible to obtain the optimal scheduling policy by using conventional CMDP methods, e.g., value iteration and policy iteration. To address this infeasibility, we exploit the Whittle index method to obtain a sub-optimal scheduling policy. In particular, we introduce a constant service charge m as the minimum service charge of the system to schedule transmission [107]. Based on this introduction, we decouple the original CMDP problem into sub-problems with a smaller state space, where each sub-problem only involves one UE. Therefore, in order to solve the problem in (2.41), we optimize the average AoI penalty of each UE individually by a Lagrange function with a multiplier m, given by

$$J^* = \min_{u(t)\in\{0,1\}} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[v\big(a(t)+d(t)(1-u(t)\delta(t))\big)+m\zeta(t)\Big]. \tag{2.42}$$

Since the AoI analysis for each UE is identical, we omit the index n from $a_n(t)$, $d_n(t)$, $\zeta_n(t)$, and $\delta_n(t)$ in (2.42) to allow easy readability. Finally, to guarantee the existence of a finite average AoI penalty, we clarify that the penalty function of AoI, $v(h)$, needs to satisfy $\sum_{k=0}^{\infty}\varepsilon^k v(k) < \infty$ and $\sum_{k=0}^{\infty}(1-\lambda)^k v(k) < \infty$. Otherwise, the average AoI penalty goes to infinity.

2.3.2 Index Policies

In this section, we analyze the Whittle index and then design the scheduling policy. Since the indexability is difficult to establish and the closed-form expressions for the Whittle index may not exist [30], we refer to find the approximated value of Whittle index to design the scheduling policy. To analyze the Whittle index, we first obtain the minimum average AoI penalty by solving (2.42). Here, we omit the time index t and present the Bellman equation as

$$J^* + \rho(a,d) = \min\{\mu_0(a,d), \mu_1(a,d)\}, \tag{2.43}$$

where

$$\mu_0(a,d) = v(a+d) + \lambda\rho(1,a+d) + (1-\lambda)\rho(a+1,d), \tag{2.44}$$

$$\begin{aligned}\mu_1(a,d) =& m+\varepsilon(v(a+d)+\lambda\rho(1,a+d)+(1-\lambda)\rho(a+1,d)) \\ &+(1-\varepsilon)\left(v(a)+\lambda\rho(1,a)+(1-\lambda)\rho(a+1,0)\right),\end{aligned} \tag{2.45}$$

and $\rho(a,d)$ is the differential cost-to-go function with $\rho(1,0)=0$. We assume that given a, $\rho(a,d)$ in (2.43) is non-decreasing with d, i.e., $\rho(a,0) \le \rho(a,1) \le \rho(a,2) \le \cdots$. Based on this, the optimal policy in (2.42) is proven to be threshold-based [108], denoted by c_D. In particular, the action of state (a,d) is to idle when $d < D_a$, and to schedule when $d > D_a$, where D_a is the threshold and satisfies $D_1 \le D_2 \le \cdots \le D_a \le \cdots$. In addition, both idle and schedule actions are equally appealing for state (a, D_a). Thus, the design of this threshold-based policy c_D is equivalent to obtain the threshold D_a. Unfortunately, all states (a,d) are deeply intertwined when $\varepsilon \neq 0$, making it extremely challenging to derive the closed-form expressions for the threshold D_a. Thus, we resort to its approximation by setting the integer threshold D_a and analyze the threshold D_a in the following Theorem.

Theorem 2.1. For the threshold-based policy c_D, the approximated integer threshold D_a satisfies

$$\begin{aligned}&\lambda\varepsilon\omega(a+D_a)+\psi(a+D_a)-\varepsilon\theta(D_1+1)\\&\quad=\frac{1}{D_1}\left(a+\frac{1}{\lambda}-1\right)\left(\frac{m}{1-\varepsilon}+\sum_{h=1}^{D_1}v(h)\right)-\sum_{h=1}^{a-1}v(h),\end{aligned}\tag{2.46}$$

for $a<D_1$, and

$$\psi(a+D_a)+\lambda\varepsilon\omega(a+D_a)=\frac{m}{1-\varepsilon}+\psi(a)+\lambda\varepsilon\omega(a),\tag{2.47}$$

for $a\geq D_1$, where $\theta(h)=\sum_{k=0}^{\infty}\varepsilon^k v(h+k)$, $\psi(h)=\sum_{k=0}^{\infty}(1-\lambda)^k v(h+k)$, and $\omega(h)=\sum_{k=1}^{\infty}\varepsilon^{k-1}\theta(h+k)$ if $\varepsilon=1-\lambda$; otherwise, $\omega(h)=(\psi(h)-\theta(h))/(1-\lambda-\varepsilon)$.

Proof: We first give the approximated expression for $\rho(a,d)$ as

$$\rho(a,d)=\begin{cases}(a+d-1)J^*-\displaystyle\sum_{h=1}^{a+d-1}v(h)\text{ , if }1\leq a+d\leq D_1,\\ \dfrac{m}{1-\varepsilon}-\dfrac{J^*}{\lambda}+\lambda\varepsilon\omega(a+d)\ \text{ if }a+d>D_1\\ -\varepsilon\theta(D_1+1)+\psi(a+d)\text{ , and }\ d\leq D_a,\\ \rho(a,D_a)+\varepsilon\left(\theta(a+d)-\theta(a+D_a)\right)\text{, otherwise.}\end{cases}\tag{2.48}$$

We then prove that (2.48) is a valid approximation to (2.43).

Since the threshold D_a can be obtained by computing $\rho(a,d)$, we analyze $\rho(a,d)$ for $d>D_a$ and $d\leq D_a$, separately.

We first analyze $\rho(a,d)$ for $d>D_a$. For the threshold-based policy c_D with the threshold D_a, the optimal action of state (a,d) is to schedule, when $d>D_a$. Then we define $\rho^{(1)}(\xi;\alpha,\sigma)\triangleq\rho(\xi,\alpha)-\rho(\xi,\sigma)$. Given a, we obtain $\rho^{(1)}(1;d,D_a)$ as

$$\begin{aligned}\rho^{(1)}(1;d,D_a)&=\varepsilon(\theta(d+1)-\theta(D_a+1))+\sum_{k=1}^{a-1}\Big(\lambda(1-\lambda)^{k-1}\big(\rho^{(1)}(k;d+a-k,D_a+a-k)\\&-\varepsilon(\theta(d+a)-\theta(D_a+a))\big)+(1-\lambda)^{a-1}\big(\rho^{(1)}(a;d,D_a)-\varepsilon\left(\theta(d+a)-\theta(D_a+a)\right)\big)\Big).\end{aligned}\tag{2.49}$$

Since (2.49) holds for any a, we obtain

$$g(1,d)=\sum_{k=1}^{a-1}\lambda(1-\lambda)^{k-1}g(k,a+d-k)+(1-\lambda)^{a-1}g(a,d),\tag{2.50}$$

where $g(a,d)=\rho^{(1)}(a;d,D_a)-\varepsilon(\theta(a+d)-\theta(a+D_a))$. We find that $g(a,d)=0$ is a valid solution to (2.50). Thus, we obtain $\rho(a,d)$ for $d > D_a$ as (2.48).

We then analyze $\rho(a,d)$ for $d \le D_a$. For the threshold-based policy c_D with the threshold D_a, the optimal action of state (a,d) is to idle, when $d < D_a$. Then, we obtain $\rho(a,d)$ as

$$\begin{aligned}\rho(a,d) &= -J^* + v(d+a) + \lambda\rho(1,d+a) + (1-\lambda)\rho(a+1,d)\\ &= -J^* + v(d+a) + \lambda\rho(1,D_1) + \varepsilon\lambda(\theta(a+d+1) - \theta(D_1+1)) + (1-\lambda)\rho(a+1,d),\end{aligned} \tag{2.51}$$

for any a, d, and k satisfying $d \le D_a$ and $a+d > D_1$. Expanding the last term recursively, we obtain

$$\begin{aligned}\rho(a,d) =& \sum_{s=0}^{k-1}(1-\lambda)^s(-J^* + v(d+a+s) + \lambda\rho(1,D_1))\\ &+\varepsilon\lambda\sum_{s=0}^{k-1}(1-\lambda)^s(\theta(a+d+1+s) - \theta(D_1+1)) + (1-\lambda)^k\rho(a+k,d)\\ =& \rho(1,D_1) - \frac{J^*}{\lambda} + \lambda\varepsilon\omega(a+d) - \varepsilon\theta(D_1+1) + \psi(a+d)\\ &+(1-\lambda)^k\big(\rho(a+k,d) - \rho(1,D_1) + \frac{J^*}{\lambda} - \lambda\varepsilon\omega(a+d+k) + \varepsilon\theta(D_1+1) - \psi(a+d+k)\big),\end{aligned} \tag{2.52}$$

From (2.52), we obtain $\rho(a,d)$ as

$$\rho(a,d) = \rho(1,D_1) - \frac{J^*}{\lambda} + \lambda\varepsilon\omega(a+d) - \varepsilon\theta(D_1+1) + \psi(a+d), \tag{2.53}$$

for $a+d \ge D_1$. In addition, we observe from (2.53) that $\rho(a_1,d_1) = \rho(a_2,d_2)$ for any a_1, a_2, $d_1 < D_{a_1}$, and $d_2 < D_{a_2}$. By combining (2.43) with $\rho(a_1,d_1) = \rho(a_2,d_2)$, we obtain

$$\rho(a,0) = J^* - v(a-1) + \rho(a-1,0) = (a-1)J^* - \sum_{h=1}^{a-1} v(h), \tag{2.54}$$

for $a \le D_1$. If D_a is an integer, we clarify that both actions for the state (a,D_a) are optimal and formulate it as

$$\mu_0(a,D_a) = \mu_1(a,D_a). \tag{2.55}$$

Based on (2.55), we obtain

$$\begin{aligned}\rho(a,D_a) &= \mu_0(a,D_a) - J^* = m - J^* + \varepsilon\mu_0(a,D_a) + (1-\varepsilon)\mu_0(a,0)\\ &= m + \varepsilon\rho(a,D_a) + (1-\varepsilon)\rho(a,0).\end{aligned} \tag{2.56}$$

Hence, we obtain $\rho(a,D_a)$ as

$$\rho(a,D_a) = \frac{m}{1-\varepsilon} + \rho(a,0). \tag{2.57}$$

Moreover, based on $\rho(1,0) = 0$, we obtain

$$\rho(1,D_1) = \frac{m}{1-\varepsilon} + \rho(1,0) = \frac{m}{1-\varepsilon}. \tag{2.58}$$

By substituting (2.58) into (2.53) and combining (2.53) with (2.54), we obtain the solution of $\rho(a,d)$ for $d < D_a$ as the first and the second cases in (2.48). Furthermore, by combining (2.48) with $\mu_0(1,D_1) = \mu_1(1,D_1)$, weexpress the optimal AoI, J^*, as a function of the threshold D_1 and the service charge m, given by

$$J^* = \frac{1}{D_1}\left(\frac{m}{1-\varepsilon} + \sum_{h=1}^{D_1} v(h)\right). \tag{2.59}$$

We then show that $\rho(a,d)$ is non-decreasing with d. Since $\rho(a,d)$ are given for three cases in (2.48), we first prove that $\rho(a,d)$ is non-decreasing with d for these three cases separately. We note that $\psi(h)$, $\theta(h)$, and $\omega(h)$ are non-decreasing functions with h, since $v(h)$ is a non-decreasing function with h. Based on the expression for $\rho(a,d)$ in (2.48), we obtain Based on the expression for $\rho(a,d)$ in (2.48), we obtain

$$\rho(a,d_2) - \rho(a,d_1) = \varepsilon(\theta(a+d_2) - \theta(a+d_1)) \geq 0, \tag{2.60}$$

if $D_a \leq d_1 \leq d_2$. Differently, if $d_1 \leq d_2 \leq D_a$ and $D_1 < a+d_1 \leq a+d_2$, we obtain

$$\rho(a,d_2) - \rho(a,d_1) = \lambda\varepsilon(\omega(a+d_2) - \omega(a+d_1)) + (\psi(a+d_2) - \psi(a+d_1)) \geq 0. \tag{2.61}$$

Based on (2.60) and (2.61), we have proved that $\rho(a,d)$ is non-decreasing with d for the second case, i.e., $a+d > D_1$ and $d \leq D_a$, and the third case, i.e., $d > D_a$, in (2.48), respectively. We then prove that $\rho(a,d)$ is non-decreasing with d for the first case, i.e., $1 \leq a+d \leq D_1$, in (2.48). Since $\mu_0(a,D_a) = \mu_1(a,D_a)$, by substituting $(a,d) = (1,D_1)$ into (2.53), we obtain

$$J^* = \lambda(\lambda\varepsilon\omega(D_1+1) - \varepsilon\theta(D_1+1) + \psi(D_1+1)) = \lambda(1-\varepsilon)\omega(D_1). \tag{2.62}$$

We note that

$$\omega(D_1) = \frac{\psi(D_1) - \theta(D_1)}{1-\lambda-\varepsilon} = \frac{\sum_{k=0}^{\infty}((1-\lambda)^k - \varepsilon^k)v(D_1+k)}{1-\lambda-\varepsilon}$$
$$\geq \frac{\sum_{k=1}^{\infty}((1-\lambda)^k - \varepsilon^k)v(D_1+1)}{1-\lambda-\varepsilon} = \frac{v(D_1+1)}{\lambda(1-\varepsilon)}. \tag{2.63}$$

Hence, we obtain

$$J^* = \lambda(1-\varepsilon)\omega(D_1) \geq v(D_1+1). \tag{2.64}$$

Based on the approximated expression for $\rho(a,d)$ in (2.48), we obtain

$$\rho(a,d_2) - \rho(a,d_1) = (d_2-d_1)J^* - \sum_{a+d_1}^{a+d_2-1} v(h) \geq 0, \tag{2.65}$$

if $a+d_1 \leq a+d_2 \leq D_1$. Hence, we have proved that $\rho(a,d)$ is non-decreasing function with d for three cases in (2.48).

We note that $\rho(a,d)$ for the first case and the second case in (2.48) are equal to each other when $a+d = D_1+1$, i.e.,

$$D_1 J^* - \sum_{h=1}^{D_1} v(h) = \frac{m}{1-\varepsilon} - \frac{J^*}{\lambda} + \lambda\varepsilon\omega(D_1+1) - \varepsilon\theta(D_1+1) + \psi(D_1+1). \tag{2.66}$$

Combining (2.61), (2.65), with (2.66), we obtain that $\rho(a,d)$ is non-decreasing for $d \leq D_a$.

Moreover, we note that

$$\rho(a,d_1) \leq \rho(a,D_a) \tag{2.67}$$

and

$$\rho(a,d_2) = \rho(a,D_a) + \varepsilon(\theta(a+d_2) - \theta(a+D_a)) \geq \rho(a,D_a), \tag{2.68}$$

when $d_1 \leq D_a \leq d_2$. Combining (2.67), (2.68), (2.60), with the fact that $\rho(a,d)$ is non-decreasing for $d \leq D_a$, we obtain that $\rho(a,d)$ is a non-decreasing function with d.

Based on the cost-to-go function $\rho(a,d)$ obtained in (2.48), we next analyze the threshold

D_a of the policy c_D. Then, when $1 \le a < D_1$, we express (2.55) as

$$\begin{aligned}
&\mu_0(a,D_a) - \mu_1(a,D_a)\\
&= -m + (1-\varepsilon)\Big(v(a+D_a) - v(a) + \lambda\rho^{(1)}(1;a+D_a,a) + (1-\lambda)\rho^{(1)}(a+1;D_a,0)\Big)\\
&= -\left(a+\frac{1}{\lambda}-1\right)J^* + \sum_{h=1}^{a-1} v(h) + \lambda\varepsilon\omega(a+D_a) - \varepsilon\theta(D_1+1) + \psi(a+D_a) = 0. \qquad (2.69)
\end{aligned}$$

By substituting (2.59) into (2.69), we obtain (2.46). It is noted that since both $\omega(h)$ and $\psi(h)$ monotonically increase with h, there is at most one possible number D_a satisfying (2.46).

When $a \ge D_1$, we express (2.55) as

$$\begin{aligned}
&\mu_0(a,D_a) - \mu_1(a,D_a)\\
&= -m + (1-\varepsilon)\big(v(a+D_a) - v(a) + \lambda\rho^{(1)}(1;a+D_a,a) + (1-\lambda)\rho^{(1)}(a+1;D_a,0)\big)\\
&= (1-\varepsilon)\left(\psi(a+D_a) - \psi(a) + \lambda\varepsilon\left(\omega(a+D_a) - \omega(a)\right)\right) - m = 0, \qquad (2.70)
\end{aligned}$$

which leads to (2.47).

■

The Whittle index is defined as the minimum auxiliary service charge to ensure that both the action of being scheduled and the action of being idle are equally appealing for the current state [107]. In other words, the Whittle index is obtained as the minimum m to obtain $d = D_a$ for the state (a,d). Based on this, we approximate the value of Whittle index in Theorem 2.2 by utilizing the approximated threshold.

Theorem 2.2. Let us consider one UE that has a packet generation probability λ and a transmission error probability ε. When this UE is in the state (a,d), its approximated Whittle index is given by

$$I_v(a,d,\lambda,\varepsilon) = \begin{cases} (1-\varepsilon)\Big(\lambda(1-\varepsilon)d\omega(d) - \sum_{h=1}^{d} v(h)\Big), & \text{if } a=1,\\ (1-\varepsilon)\Big(\lambda(1-\varepsilon)D_1\omega(D_1) - \sum_{h=1}^{D_1} v(h)\Big), & \text{if } 2 \le a \le D_1,\\ (1-\varepsilon)\big(\psi(a+d) - \psi(a) + \lambda\varepsilon(\omega(a+d) - \omega(a))\big), & \text{if } a > D_1, \end{cases} \qquad (2.71)$$

where D_1 is the minimum positive integer number satisfying

$$\varepsilon\theta(D_1+1) + \lambda(1-\varepsilon)\left(a+\frac{1}{\lambda}-1\right)\omega(D_1) \ge \lambda\varepsilon\omega(a+d) + \psi(a+d) + \sum_{h=1}^{a-1} v(h). \qquad (2.72)$$

Proof: When $a \le D_1$, we combine (2.59) and (2.62) as

$$I_v(1,d,\lambda,\varepsilon) = (1-\varepsilon)\left(\lambda(1-\varepsilon)D_1\omega(D_1) - \sum_{h=1}^{D_1} v(h)\right). \tag{2.73}$$

We note that $D_1 = d$ holds for $a = 1$. Otherwise, by substituting $D_a = d$ into (2.46), the threshold D_1 is the minimum positive number satisfying (2.72).

When $a > D_1$, we obtain the Whittle index from (2.47) as

$$I_v(a,d,\lambda,\varepsilon) = (1-\varepsilon)\big(\psi(a+d) - \psi(a) + \lambda\varepsilon\left(\omega(a+d) - \omega(a)\right)\big). \tag{2.74}$$

With (2.73) and (2.74), the Whittle index is obtained as (2.71). ■

We clarify that the derived approximated Whittle index is smaller than its exact value since the approximate threshold D_a derived in Theorem 2.1 is higher than the exact threshold D_a. We remark that when considering the special case with a deterministic packet generation model, i.e., $\lambda = 1$, the Whittle index in (2.71) coincides with the results in [38], which gives the closed-form expressions for the Whittle index. Furthermore, when considering the average AoI penalty being given by the average AoI, i.e., $v(h) = h$, the derived Whittle index in (2.71) matches the result in [30]. Therefore, the derived approximated Whittle index in (2.71) is a general approximation for stochastic packet generation models and any penalty function of AoI, and it is the closed-form expression if $\lambda = 1$.

Based on the derived Whittle index in (2.71), we propose the near optimal transmission scheduling policy to minimize the average AoI penalty. In particular, for each time slot, the BS schedules the UE with the highest values of Whittle index for transmission. This is because for the UE with a larger Whittle index, the transmission of its packet makes more contributions to reducing the average AoI penalty in the system.

2.3.3 Numerical Result and Discussion

In this section, we present numerical results to demonstrate the effectiveness of our proposed Whittle index based scheduling policy in Section 2.3.2. For comparison, we first compare our scheduling policy with the optimal scheduling policy, where the optimal scheduling policy is obtained by utilizing the relative value iteration to solve the CMDP problem in (2.41) for a two-UE scenario. We then employ the on-demand Whittle Index policy [30] and the age greedy scheduling policy [38] as the benchmark policies.

In Fig. 2.5, we compare the proposed scheduling policy with the optimal scheduling policy for two users with an AoI violation function as the penalty function such that $v(h) = 1$ if $h \ge 6$; otherwise, $v(h) = 0$. It turns out that the designed scheduling policy almost achieves the minimum average AoI penalty, which shows the optimality of our designed scheduling

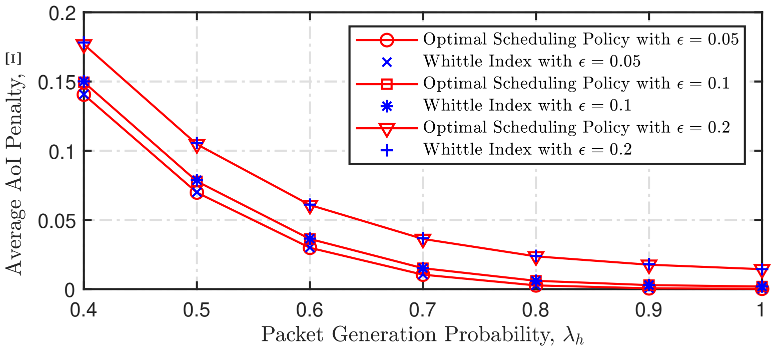


Figure 2.5: The average AoI penalty for two users under the proposed Whittle index scheduling policy and an optimal scheduling policy.

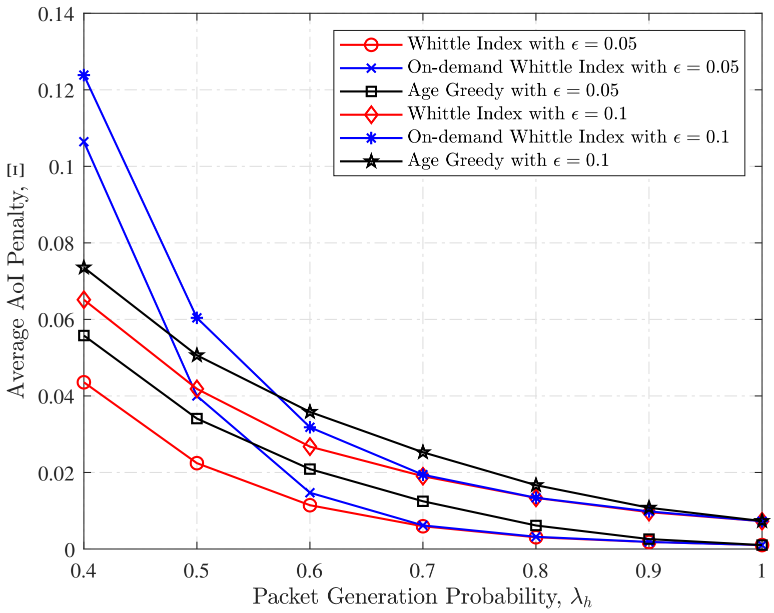


Figure 2.6: The average AoI penalty versus the packet generation probability with $N = 6$.

policy.

Fig. 2.6. plots the average AoI penalty versus the packet generation probability, λ, for the homogeneous case where all UEs share the same packet generation pattern and the same transmission error probability, i.e, $\lambda_n = \lambda$, and $\varepsilon_n = \varepsilon$ for $n = 1, 2, \cdots, N$. We employ an AoI violation function as the penalty function such that $v(h) = 1$ if $h \geq 10$; otherwise, $v(h) = 0$. We observe that for all ε, the designed scheduling policy achieves a lower average AoI penalty than the benchmark policies, especially for small λ_h, which shows the advantage of the designed scheduling policy over existing policies. This is due to the fact that the impact of packet freshness on the average AoI penalty increases when λ decreases and our proposed policy

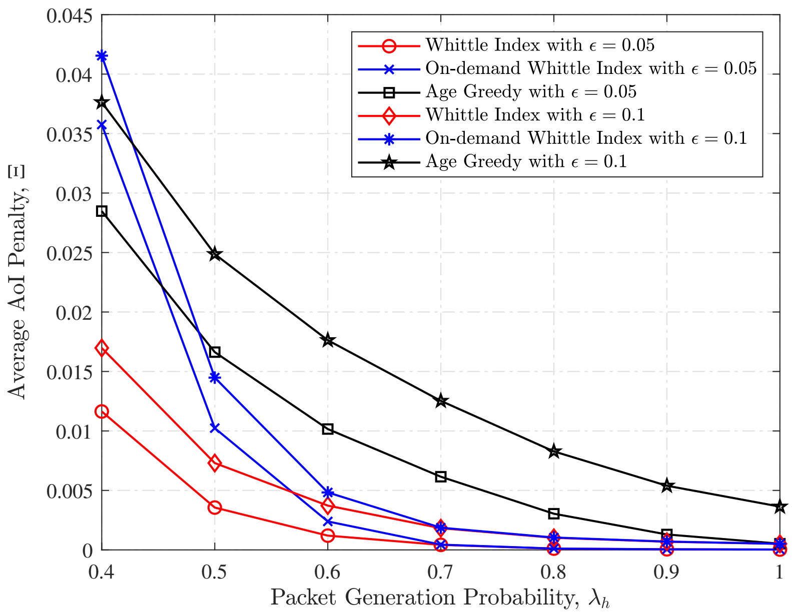


Figure 2.7: The average AoI penalty versus the packet generation probability with $N = 6$.

design is forward-looking instead of myopic, i.e., it evaluates the long term payoff.

Fig. 2.7. plots the average AoI penalty versus the packet generation probability, λ, for the heterogeneous case where half UEs have the same penalty function such that $v(h) = 1$ if $h \geq 10$; otherwise, $v(h) = 0$. The other half UEs have another penalty function such that $v(h) = 1$ if $h \geq 15$; otherwise, $v(h) = 0$. Given different penalty functions among UEs, we observe that compared to the benchmark policies, our proposed policy achieves a larger AoI performance gain for the heterogeneous case than for the homogeneous case. This is due to the fact that UEs have different dissatisfaction levels of data staleness, which results in different impacts of UEs' successful packet transmission on the reduction in the average AoI penalty. Importantly, such impacts are addressed in our proposed policy.

2.3.4 Summary

In this section, we considered a multi-user system where a BS generates time-sensitive packets and transmits them to UEs under an unreliable channel. We employed the general AoI penalty as the metric to characterize the dissatisfaction level of receivers on transmitted packets. By introducing a constant service charge, we derived the approximated Whittle index. Based on this expression, we proposed a Whittle index based scheduling policy. Under this proposed scheduling policy, Whittle index reveals the value of each packet and the BS intends to send the most valuable packet to reduce the average AoI penalty. Using simulations, we showed the performance gain of our proposed policy.

Chapter 3

AoI Analysis and Optimization of MEC Systems

3.1 Introduction

To find the solution to the key problem: "How many computational tasks need to be offloaded in a multi-user MEC system such that its AoI performance is optimized?", we investigate the impact of computing schemes on the AoI performance in an MEC system in this chapter. In particular, we analyze the average AoI and the average PAoI of a multi-user MEC system where a BS generates and transmits computation-intensive packets to UEs. In this MEC system, we focus on three computing schemes: (i) The local computing scheme where all computational tasks are computed by the local server at the UE, (ii) The edge computing scheme where all computational tasks are computed by the edge server at the BS, and (iii) The partial computing scheme where computational tasks are partially allocated at the edge server and the rest are computed by the local server. Considering exponentially distributed transmission time and computation time and adopting the FCFS queuing policy, we derive closed-form expressions for the average AoI and average PAoI. Due to the complex expression for the average AoI, we further derive a simple expression for the upper bound and lower bound on the average AoI, which allow us to explicitly examine the dependence of the optimal offloading decision on the MEC system parameters. Aided by simulation results, we verify our analysis and illustrate the impact of system parameters on the offloading strategy design.

This chapter is organized as follows. In Section 3.2, the system model and the average AoI and PAoI of the considered MEC system are presented. The closed-form expressions for the average AoI and PAoI of three schemes are derived in Section 3.3. In Section 3.4, the upper bound and the lower bound on the average AoI and the optimal offloading ratio are presented. The numerical results are discussed in Section 3.5. The summary is concluded in Section 3.6.

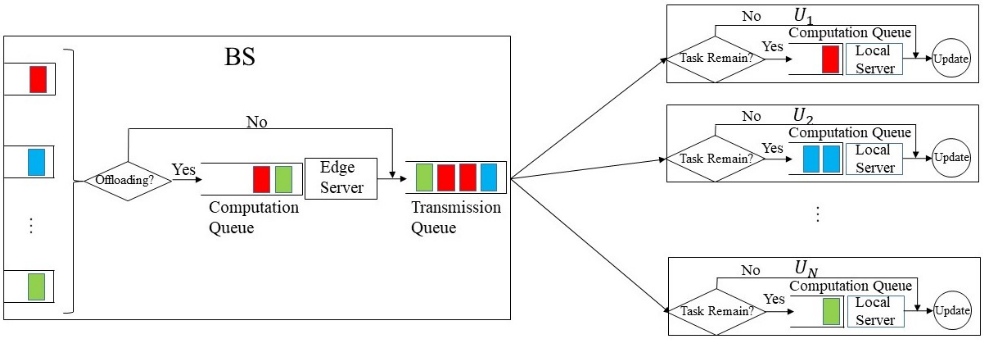


Figure 3.1: Illustration of our considered MEC system where the BS transmits computation-intensive packets to N UEs.

3.2 System Model and Average AoI

We consider a downlink system, as depicted in Fig. 3.1, where the BS transmits time-sensitive packets to N UEs. We denote the nth UE by U_n, where $n = 1, 2, \cdots, N$. In this system, the BS generates the packet of U_n according to a Poisson process[1] with the rate λ_n[2]. We assume that packets are generated with indices (e.g., using the first several bits of each packet) to associate with UEs and the BS transmits packets to their corresponding UEs based on the indices. To ensure the freshness of packets, we consider an MEC system, where a portion of total computational tasks of a packet can be processed at the local server and the remaining part is processed at the edge server. Such separate processing is reflected in practical applications, such as the virus scan application and the file compression application [109], where the computational tasks of packets are divided into several independent subsets that can be processed individually. In this MEC system, we introduce three computing schemes, namely, the local computing scheme, the edge computing scheme, and the partial computing scheme.

3.2.1 Computing Schemes

Depending on which server computes the computational tasks of packets, we introduce three computing schemes as follows:

1. *Local Computing*: In this scheme, the BS directly transmits the generated packet to the UE for local computing. In particular, when a packet of U_n is generated by the BS, denoted by P_n, it waits to be transmitted in the transmission queue at the BS. After being

[1] We assume that the packet generation processes among UEs are independent but not identical.

[2] In practical applications, e.g., a vehicular network, the BS can generate traffic information for vehicles. In this vehicular network, the BS has two options of traffic information transmission and computation. One option is that the BS transmits the traffic information to each vehicle and each vehicle computes the action command according to the received traffic information. The other option is that the BS computes the action command and sends it to each vehicle.

transmitted to U_n, P_n arrives at the computation queue at the local server at U_n. Finally, the local server completes the computation of P_n and U_n obtains the computational result of P_n.

2. *Edge Computing*: In this scheme, the edge server at the BS performs the computation of the packet and transmits the computational result to the UE. In particular, the generated packet P_n arrives at the computation queue at the edge server. After the computation is completed by the edge server, the computational result of P_n arrives at the transmission queue and is transmitted to U_n.

3. *Partial Computing*: In this scheme, the edge server partially computes the packet and then the BS transmits the intermediate computational result to the UE for the remaining computation. Specifically, the generated packet P_n first arrives at the computation queue at the edge server for a partial computation and the intermediate computational result of P_n arrives at the transmission queue. When U_n receives the intermediate computational result of P_n, its local server performs the remaining computation to obtain the computational result of P_n.

In these three computing schemes, we assume that the packets in the queues are served by using the FCFS queuing policy and each queue has infinite buffer size to store the packet[3]. We then assume that both the computation time and the transmission time of a packet follow exponential distributions[4], which are commonly used in the AoI analysis, e.g., [18, 53, 110]. On one hand, the packet transmission time depends on wireless channels, which may or may not incorporate retransmission and backoff. It has been shown that such complex effects are accurately characterized by an exponential distributed transmission time in [18]. On the other hand, the packet computation time depends on the complexity of computational tasks. For example, [110] showed that an exponential distributed computation time is accurate to characterize the different complexity of computational tasks in face recognition. Then, the probability density functions (PDFs) of the computation time of a packet at the edge server, S_B, the transmission time of a packet, S_D, and the computation time of a packet at the local server at U_n, S_{U_n}, are given by

$$f_{S_B}(t) = \exp(-\mu_B t), \tag{3.1}$$

[3]The FCFS queuing policy is easier to be implemented in practice, comparing with the LCFS queuing policy. In addition, the LCFS queuing policy may lead to unfairness and poor performance for some UEs in multi-user systems.

[4]We clarify that our analytical framework accommodates other distributions, if we use the PDFs of other distributions to replace the PDFs of the computation time of a packet at the edge server, S_B, the transmission time of a packet, S_D, and the computation time of a packet at the local server at U_n, S_{U_n}, i.e., $f_{S_B}(t)$, $f_{S_D}(t)$, and $f_{S_{U_n}}(t)$.

$$f_{S_D}(t) = \exp(-\mu_D t), \tag{3.2}$$

and

$$f_{S_{U_n}}(t) = \exp(-\mu_n t), \tag{3.3}$$

where μ_B is the computation rate of the edge server, μ_D is the transmission rate, and μ_n is the computation rate of the local server at U_n. Here, μ_B indicates the average number of packets served per unit time at the edge server in the edge computing scheme, μ_n indicates the average number of packets served per unit time at the local server at U_n in the local computing scheme, and μ_D indicates the average number of packets transmitted per unit time from the BS to UEs[5].

In order to fairly compare the AoI performance of these three schemes, we assume that these three computing schemes share the same transmission rate and computation rate at each server. For the partial computing scheme, we adopt a linear computation partitioning model introduced in [109]. In particular, we denote p as the offloading ratio, which represents the percentage of computational tasks of each packet computed by the edge server, where $0 \leq p \leq 1$. Moreover, the effective computation rate of the edge server is denoted by μ'_B and the effective computation rate of the local server at U_n is denoted by μ'_n. Here, μ'_B and μ'_n indicate the average number of packets served per unit time at the edge server and the local server at U_n, respectively, which depend on not only the performance of servers but the proportion of tasks that need to be computed in each packet at the server. Thus, we have $\mu'_B = \frac{\mu_B}{p}$ and $\mu'_n = \frac{\mu_n}{1-p}$. We clarify that the local computing scheme is considered as a special case of the partial computing scheme where the offloading ratio is 0, i.e., $p = 0$. In this scheme, the serving time of packets at the edge server is 0. In addition, the edge computing scheme is considered as another special case of the partial computing scheme where the offloading ratio is 1, i.e., $p = 1$. In this scheme, the serving time of packets at the local server at U_n is 0.

3.2.2 Average AoI and PAoI

In this subsection, we present the general expression for the average AoI and PAoI of the considered MEC system. Recall that our target is to analyze the AoI performance of the considered MEC system. Without loss of generality, we arbitrarily select one UE, U_n, and analyze its average AoI, Δ_n, and average PAoI, Ω_n. We will define the average AoI and the average PAoI of the system at the end of this section based on the individual UE's AoI. Fig. 3.2 plots a sample variation of AoI for U_n, $\Delta_n(t)$, as a function of t. We assume that the observation begins at

[5]We assume that the information bits of the original packet, the partial computed packet, and the computed packet are same, such that the transmission rate of these packets are same.

$t = 0$, where the AoI is $\Delta_n(0)$. From Fig. 3.2, we express the AoI of U_n at time t as

$$\Delta_n(t) = t - u_n(t), \tag{3.4}$$

where $u_n(t)$ is the generation time of the last received computed packet of U_n at time t. Then the time-average AoI of U_n over the observation time interval $(0, \tau)$ can be calculated as

$$\Delta_n = \frac{1}{\tau}\int_0^{\tau} \Delta_n(t)\mathrm{d}t. \tag{3.5}$$

We denote $P_{n,j}$ as the jth packet generated after time $t = 0$ of U_n, $j = 1, 2, \cdots$. We then denote Y_j as the time interval from the generation time of $P_{n,j-1}$ to the generation time of $P_{n,j}$ and denote T_j as the time interval from the generation time of $P_{n,j}$ to the time that U_n obtains the computational result of $P_{n,j}$. The value of age in the peak is denoted by A_j. Therefore, we obtain

$$Y_j = t_j - t_{j-1}, \tag{3.6}$$

$$T_j = t'_j - t_j, \tag{3.7}$$

and

$$A_j = Y_j + T_j, \tag{3.8}$$

where t_j is the generation time of $P_{n,j}$ and t'_j is the time that U_n obtains the computational result of $P_{n,j}$. According to [12], the average AoI of U_n is calculated as

$$\Delta_n = \frac{\mathbb{E}[Y_j^2] + 2\mathbb{E}[Y_j T_j]}{2\mathbb{E}[Y_j]}. \tag{3.9}$$

We keep the index in (3.9) to reflect the correlation between Y_j and T_j. As the BS generates the packet of U_n according to a Poisson process with the rate λ_n, we obtain $\mathbb{E}[Y_j^2] = \frac{2}{\lambda_n^2}$ and $\mathbb{E}[Y_j] = \frac{1}{\lambda_n}$, which leads to

$$\Delta_n = \frac{1}{\lambda_n} + \lambda_n \mathbb{E}[Y_j T_j]. \tag{3.10}$$

For the average PAoI, we calculate it as

$$\Omega_n = \lim_{m\to\infty} \frac{1}{m}\sum_{j=1}^{m} A_j = \frac{1}{\lambda_n} + \mathbb{E}[T_j]. \tag{3.11}$$

We denote $t_{j,B}$ as the time that $P_{n,j}$ arrives at the transmission queue and $t_{j,D}$ as the time that $P_{n,j}$ arrives at the computation queue at the local server at U_n. We then denote $X_{j,B}$, $X_{j,D}$, and $X_{j,U}$ as the queuing delay of $P_{n,j}$ in the computation queue at the edge server, the queuing delay of $P_{n,j}$ in the transmission queue, and the queuing delay of $P_{n,j}$ in the computation queue at the local server at U_n, respectively. They are expressed as

$$X_{j,B} = t_{j,B} - t_j, \tag{3.12}$$

$$X_{j,D} = t_{j,D} - t_{j,B}, \tag{3.13}$$

and

$$X_{j,U} = t'_j - t_{j,D}, \tag{3.14}$$

respectively. Then, we rewrite T_j as

$$T_j = X_{j,B} + X_{j,D} + X_{j,U}. \tag{3.15}$$

We now denote $W_{j,B}$, $W_{j,D}$, and $W_{j,U}$ as the waiting time of $P_{n,j}$ in the computation queue at the edge server, the waiting time of $P_{n,j}$ in the transmission queue, and the waiting time of $P_{n,j}$ in the computation queue at the local server at U_n, respectively. Next, we denote $S_{j,B}$, $S_{j,D}$, and $S_{j,U}$ as the computation time of $P_{n,j}$ at the edge server, the transmission time of $P_{n,j}$, and the computation time of $P_{n,j}$ at the local server at U_n, respectively. Note that the queuing delay of $P_{n,j}$ in each queue is the summation of the waiting time and the serving time of $P_{n,j}$ in this queue. Thus, we rewrite $X_{j,B}$, $X_{j,D}$, and $X_{j,U}$ as

$$X_{j,B} = W_{j,B} + S_{j,B}, \tag{3.16}$$

$$X_{j,D} = W_{j,D} + S_{j,D}, \tag{3.17}$$

and

$$X_{j,U} = W_{j,U} + S_{j,U}, \tag{3.18}$$

respectively. By substituting (3.16), (3.17), and (3.18) into (3.15), we obtain

$$T_j = W_{j,B} + S_{j,B} + W_{j,D} + S_{j,D} + W_{j,U} + S_{j,U}. \tag{3.19}$$

We note that the packet generation is independent of the packet transmission and computation in each queue. By substituting (3.19) into (3.10) and (3.11), and using the fact that $S_{j,B}$, $S_{j,D}$,

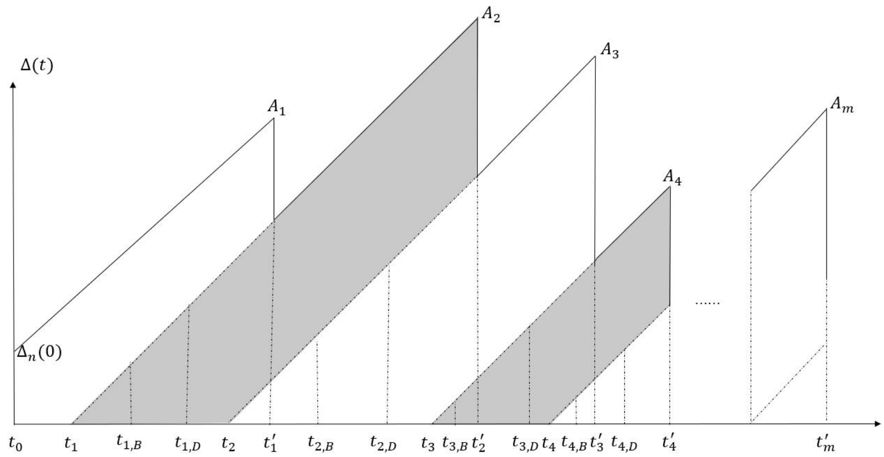


Figure 3.2: The AoI variation of the selected UE, U_n.

and $S_{j,U}$ are independent of Y_j, we obtain

$$\Delta_n = \frac{1}{\lambda_n} + \mathbb{E}[S_{j,B}] + \mathbb{E}[S_{j,D}] + \mathbb{E}[S_{j,U}] + \lambda_n \left(\mathbb{E}[Y_j W_{j,B}] + \mathbb{E}[Y_j W_{j,D}] + \mathbb{E}[Y_j W_{j,U}]\right), \quad (3.20)$$

and

$$\Omega_n = \frac{1}{\lambda_n} + \mathbb{E}[S_{j,B}] + \mathbb{E}[S_{j,D}] + \mathbb{E}[S_{j,U}] + \mathbb{E}[W_{j,B}] + \mathbb{E}[W_{j,D}] + \mathbb{E}[W_{j,U}]. \quad (3.21)$$

By averaging Δ_n over all UEs, we obtain the average AoI and PAoI of the MEC system as

$$\Delta = \frac{1}{N}\sum_{n=1}^{N}\Delta_n, \quad (3.22)$$

and

$$\Omega = \frac{1}{N}\sum_{n=1}^{N}\Omega_n, \quad (3.23)$$

respectively. It is worthwhile noting that, when one of the expectations of $W_{j,B}$, $W_{j,D}$, and $W_{j,U}$ in (3.20) becomes infinity, the average AoI and the average PAoI of U_n go to infinity. In this case, the MEC system is considered to be unstable. Thus, in order to ensure system stability, we assume that the arrival rate of the packet is lower than its serving rate in each queue.

3.3 Characterizing Average AoI and PAoI

In this section, we derive the closed-form expressions for the average AoI and PAoI of three computing schemes. We first derive the closed-form expression for the average AoI and PAoI of the partial computing scheme, since both the local computing scheme and the edge computing scheme are special cases of the partial computing scheme. Based on that, we then obtain the closed-form expressions for the average AoI and PAoI of the local computing scheme and the edge computing scheme, respectively.

Theorem 3.1. In the partial computing scheme with the offloading ratio p, the closed-form expression for the average AoI of U_n is derived as

$$\begin{aligned}\Delta_{n,p} =&\frac{1}{\lambda_n}+\frac{1}{\mu'_B}+\frac{1}{\mu_D}+\frac{1}{\mu'_n}+\frac{\lambda_{-n}}{\mu'_B(\mu'_B-\lambda_{-n})}+\frac{\lambda_n^2\lambda_{-n}}{\mu'_B(\mu'_B-\lambda_{-n})^3}\\&+\frac{\lambda_n^2}{(\mu'_B-\lambda)(\mu'_B-\lambda_{-n})^2}+\lambda_n(\Phi_{p,B_{j,D}}+\Phi_{p,L_{j,D}}+\Phi_{p,B_{j,U}}+\Phi_{p,L_{j,U}}),\end{aligned} \tag{3.24}$$

and the closed-form expression for the average PAoI of U_n is derived as

$$\Omega_{n,p} = \frac{1}{\lambda_n}+\frac{1}{\mu'_B-\lambda}+\frac{1}{\mu_D-\lambda}+\frac{1}{\mu'_n-\lambda_n}, \tag{3.25}$$

where $\lambda=\sum_{n=1}^{N}\lambda_n$ and $\lambda_{-n}=\lambda-\lambda_n$. Here, $\Phi_{p,B_{j,D}}$, $\Phi_{p,L_{j,D}}$, $\Phi_{p,B_{j,U}}$, and $\Phi_{p,L_{j,U}}$ are given by

$$\begin{aligned}\Phi_{p,B_{j,D}} =& \frac{\lambda_n(\mu'_B+\mu_D-\lambda)}{\mu'_B(\mu_D-\lambda)(\mu'_B+\mu_D-\lambda-\lambda_{-n})^2}+\frac{\lambda_n(\mu'_B-\lambda)(\mu'_B-\lambda_{-n})\left(\frac{1}{(\mu_D-\lambda_{-n})^2}-\frac{1}{(\mu'_B-\lambda_{-n})^2}\right)}{(\mu'_B-\mu_D)(\mu_D-\lambda)(\mu_D+\mu'_B-\lambda-\lambda_{-n})}\\&+\frac{\lambda_n\lambda_{-n}(\mu'_B-\lambda)(\mu'_B-\lambda_{-n})\left(\frac{2}{(\mu_D-\lambda_{-n})^3}-\frac{2}{(\mu'_B-\lambda_{-n})^3}\right)}{\mu_D(\mu'_B-\mu_D)(\mu'_B+\mu_D-\lambda-\lambda_{-n})}+\frac{2\lambda_n\lambda_{-n}(\mu_D+\mu'_B-\lambda)}{\mu'_B\mu_D(\mu_D+\mu'_B-\lambda-\lambda_{-n})^3},\end{aligned} \tag{3.26}$$

$$\Phi_{p,L_{j,D}} = \frac{\lambda_{-n}}{\mu_D(\mu_D-\lambda_{-n})}\left(\frac{1}{\lambda_n}-\frac{\lambda_n(\mu'_B+\mu_D-\lambda)}{\mu'_B(\mu'_B+\mu_D-\lambda-\lambda_{-n})^2}-\frac{\lambda_n(\mu'_B-\lambda)\left(\frac{1}{\mu'_B-\lambda_{-n}}-\frac{\mu'_B-\lambda_{-n}}{(\mu_D-\lambda_{-n})^2}\right)}{(\mu_D-\mu'_B)(\mu'_B+\mu_D-\lambda-\lambda_{-n})}\right), \tag{3.27}$$

$$\begin{aligned}\Phi_{p,B_{j,U}} =& \frac{\lambda_n(\mu'_B-\lambda)(\mu'_B-\lambda_{-n})(\mu_D+\mu'_n-\lambda_n)}{\mu_D(\mu_D-\lambda+\mu'_n)^2(\mu'_n-\lambda_n)(\mu'_B-\mu_D)(\mu'_B+\mu_D-\lambda-\lambda_{-n})}\\&-\frac{\lambda_n\mu_D(\mu'_B-\lambda)(\mu'_B-\lambda_{-n})(\mu_D+\mu'_n-\lambda_n)}{(\mu'_n-\lambda_n)(\mu'_B-\mu_D)(\mu'_B+\mu_D-\lambda-\lambda_{-n})((\mu_D+\mu'_n)\mu'_B-\lambda\mu_D+(\mu_D-\mu'_B)\lambda_n)^2}\\&+\frac{\lambda_n\mu'_B\mu_D\frac{(\mu_D+\mu'_n-\lambda_n)(\mu'_B+\mu_D-\lambda)}{\mu'_n-\lambda_n}}{((\mu'_B+\mu_D-\lambda)(\mu_D+\mu'_n-\lambda_n)(\mu'_B-\lambda_{-n})+\lambda_{-n}(\mu_D^2+(2\mu'_n-2\lambda_n-\lambda)\mu_D+(\mu'_B-2\lambda)(\mu'_n-\lambda_n)))^2},\end{aligned} \tag{3.28}$$

and

$$\Phi_{p,L_{j,U}} = \frac{\lambda_n(\mu_D-\lambda)(\mu_D-\lambda_{-n})}{\mu'_B(\mu'_n-\lambda_n)(\mu_D-\mu'_n-\lambda_{-n})(\mu_D+\mu'_n-\lambda)}\left(\frac{\mu'_B+\mu'_n-\lambda_n}{(\mu'_B+\mu'_n-\lambda)^2}-\frac{\mu'_B+\mu_D-\lambda}{(\mu'_B+\mu_D-\lambda-\lambda_{-n})^2}\right)$$
$$+\frac{\lambda_n(\mu'_B-\lambda)(\mu'_B-\lambda_{-n})(\mu_D-\lambda)(\mu_D-\lambda_{-n})\left(\frac{1}{\mu'^2_n}-\frac{1}{(\mu'_B-\lambda_{-n})^2}\right)}{(\mu'_n-\lambda_n)(\mu_D-\mu'_n-\lambda_{-n})(\mu_D+\mu'_n-\lambda)(\mu'_B+\mu'_n-\lambda)(\mu'_B-\mu'_n-\lambda_{-n})}$$
$$-\frac{\lambda_n(\mu'_B-\lambda)(\mu'_B-\lambda_{-n})(\mu_D-\lambda)(\mu_D-\lambda_{-n})\left(\frac{1}{(\mu_D-\lambda_{-n})^2}-\frac{1}{(\mu'_B-\lambda_{-n})^2}\right)}{(\mu'_n-\lambda_n)(\mu_D-\mu'_n-\lambda_{-n})(\mu'_B-\mu_D)(\mu'_B+\mu_D-\lambda-\lambda_{-n})}, \tag{3.29}$$

respectively.

Proof: See Appendix A.1. ■

From Theorem 3.1, we find that the average PAoI has a concise expression than the average AoI[6]. Based on (3.24), we obtain the average AoI and the average PAoI in the local computing and the edge computing scheme by setting $p = 0$ and $p = 1$, which are given in the following two corollaries, respectively.

Corollary 3.1. In the local computing scheme, the closed-form expression for the average AoI of U_n is derived as

$$\Delta_{n,l} = \frac{1}{\lambda_n}+\frac{1}{\mu_D}+\frac{1}{\mu_n}+\frac{\lambda_{-n}}{\mu_D(\mu_D-\lambda_{-n})}+\frac{\lambda_n^2\lambda_{-n}}{\mu_D(\mu_D-\lambda_{-n})^3}+\frac{\lambda_n^2}{(\mu_D-\lambda)(\mu_D-\lambda_{-n})^2}$$
$$+\frac{\lambda_n^2(\mu_D+\mu_n-\lambda_n)}{\mu_D(\mu_n-\lambda_n)(\mu_D+\mu_n-\lambda)^2}+\frac{\lambda_n^2(\mu_D-\lambda)(\mu_D+\mu_n-\lambda_{-n})}{\mu_n^2(\mu_D-\lambda_{-n})(\mu_n-\lambda_n)(\mu_D+\mu_n-\lambda)}, \tag{3.30}$$

and the closed-form expression for the average PAoI of U_n is derived as

$$\Omega_{n,l} = \frac{1}{\lambda_n}+\frac{1}{\mu_D-\lambda}+\frac{1}{\mu_n-\lambda_n}. \tag{3.31}$$

Proof: In this scheme, all the tasks in each packet of U_n are computed by the local server at U_n. It can be regarded as a partial computing scheme with $p = 0$, where the computation rate of the local server at U_n is given as $\mu'_n = \mu_n$ and the computation rate of the edge server is given as $\mu'_B \to \infty$. Therefore, we obtain the closed-form expression for the average AoI of U_n in the local computing scheme, $\Delta_{n,l}$, and the average PAoI of U_n in the local computing scheme, $\Omega_{n,l}$, as

$$\Delta_{n,l} = \Delta_{n,p}\big|_{\mu'_B\to\infty,\mu'_n=\mu_n}, \tag{3.32}$$

[6]Recently, the shifted exponential distributed computation time is evaluated in [111], since the task computation requires a minimum computation time. We note that the shifted exponential distributed computation time is a fixed computation time with the original exponential distributed computation time. Hence, the impact of this distribution on both the average AoI and the average PAoI is a constant queuing delay in computation queues at both the edge server and the local server.

and

$$\Omega_{n,l} = \Omega_{n,p}\big|_{\mu_B' \to \infty, \mu_n' = \mu_n}, \tag{3.33}$$

respectively, which are given in (3.30) and (3.31). ■

Corollary 3.2. In the edge computing scheme, the closed-form expression for the average AoI of U_n is derived as

$$\begin{aligned}\Delta_{n,e} =& \frac{1}{\lambda_n} + \frac{1}{\mu_B} + \frac{1}{\mu_D} + \frac{\lambda_{-n}}{\mu_B(\mu_B - \lambda_{-n})} + \frac{\lambda_n^2 \lambda_{-n}}{\mu_B(\mu_B - \lambda_{-n})^3} \\ &+ \frac{\lambda_n^2}{(\mu_B - \lambda)(\mu_B - \lambda_{-n})^2} + \lambda_n(\Phi_{e,B_{j,D}} + \Phi_{e,L_{j,D}}),\end{aligned} \tag{3.34}$$

and the closed-form expression for the average PAoI of U_n is derived as

$$\Omega_{n,l} = \frac{1}{\lambda_n} + \frac{1}{\mu_B - \lambda} + \frac{1}{\mu_D - \lambda}, \tag{3.35}$$

where $\Phi_{e,B_{j,D}}$ and $\Phi_{e,L_{j,D}}$ are given by

$$\begin{aligned}\Phi_{e,B_{j,D}} =& \frac{\lambda_n(\mu_B + \mu_D - \lambda)}{\mu_B(\mu_D - \lambda)(\mu_B + \mu_D - \lambda - \lambda_{-n})^2} + \frac{\lambda_n(\mu_B - \lambda)(\mu_B - \lambda_{-n})\left(\frac{1}{(\mu_D - \lambda_{-n})^2} - \frac{1}{(\mu_B - \lambda_{-n})^2}\right)}{(\mu_B - \mu_D)(\mu_D - \lambda)(\mu_D + \mu_B - \lambda - \lambda_{-n})} \\ &+ \frac{\lambda_n \lambda_{-n}(\mu_B - \lambda)(\mu_B - \lambda_{-n})\left(\frac{2}{(\mu_D - \lambda_{-n})^3} - \frac{2}{(\mu_B - \lambda_{-n})^3}\right)}{\mu_D(\mu_B - \mu_D)(\mu_B + \mu_D - \lambda - \lambda_{-n})} + \frac{2\lambda_n \lambda_{-n}(\mu_D + \mu_B - \lambda)}{\mu_B \mu_D(\mu_D + \mu_B - \lambda - \lambda_{-n})^3}\end{aligned} \tag{3.36}$$

and

$$\Phi_{e,L_{j,D}} = \frac{\lambda_{-n}}{\mu_D(\mu_D - \lambda_{-n})}\left(\frac{1}{\lambda_n} - \frac{\lambda_n(\mu_B + \mu_D - \lambda)}{\mu_B(\mu_B + \mu_D - \lambda - \lambda_{-n})^2} - \frac{\lambda_n(\mu_B - \lambda)\left(\frac{1}{\mu_B - \lambda_{-n}} - \frac{\mu_B - \lambda_{-n}}{(\mu_D - \lambda_{-n})^2}\right)}{(\mu_D - \mu_B)(\mu_B + \mu_D - \lambda - \lambda_{-n})}\right), \tag{3.37}$$

respectively.

Proof: In this scheme, all the tasks in each packet are computed by the edge server. It can be regarded as a partial computing scheme with $p = 1$, where the computation rate of the local server at U_n is given as $\mu_n' \to \infty$ and the computation rate of the edge server is given as $\mu_B' = \mu_B$. Therefore, we obtain the closed-form expression for the average AoI of U_n in the edge computing scheme, $\Delta_{n,e}$, and the average PAoI of U_n in the edge computing scheme, $\Omega_{n,e}$, as

$$\Delta_{n,e} = \Delta_{n,p}\big|_{\mu_B' = \mu_B, \mu_n' \to \infty}, \tag{3.38}$$

and

$$\Omega_{n,e} = \Omega_{n,p}\big|_{\mu'_B=\mu_B,\mu'_n\to\infty}, \tag{3.39}$$

respectively, which are given in (3.34) and (3.35). ■

From Theorem 3.1, Corollary 3.1, and Corollary 3.2, we find that the average AoI has a more complex closed-form expression than the average PAoI. Due to this complexity, it is difficult to derive the closed-form expression for the optimal offloading ratio to minimize the average AoI. Comparing with the average AoI, we find that the expression for the average PAoI is less complex and it is feasible to obtain the closed-form expression for the optimal offloading ratio to minimize the average PAoI. Moreover, we note that the results in Theorem 3.1 can be used to obtain the results for a single UE system by setting $\lambda = \lambda_n$ and $\lambda_{-n} = 0$.

3.4 AoI Performance Optimization

In this section, we consider an MEC system with homogeneous UEs, where all UEs share the same packet generation rate λ_h and the same computation rate of the local server $\mu'_h = \frac{\mu_h}{1-p}$, i.e., $\lambda_n = \lambda_h$ and $\mu_n = \mu_h$, $\forall\, n$. In this homogeneous scenario, due to the complex expression for the average AoI given in (3.24), it is hard to derive the closed-form expression for the optimal offloading ratio to minimize the average AoI.

To find the closed-form expression for the optimal offloading ratio to minimize the average AoI, we first derive the simplified upper bound and lower bound on the average AoI.

Theorem 3.2. In the MEC system under the aforementioned assumptions, the simple form lower bound and upper bound on the average AoI are given as

$$\Delta_{low} = \Omega - \Delta_{gap}, \tag{3.40}$$

and

$$\Delta_{up} = \Omega, \tag{3.41}$$

respectively, where $\Omega = \frac{p}{\mu_B - p\lambda} + \frac{1}{\mu_D - \lambda} + \frac{1-p}{\mu_h - (1-p)\lambda_h}$ and

$$\Delta_{gap} = \frac{\lambda_h}{(\mu'_B - \lambda_{-h})^2} + \frac{\lambda_h}{(\mu_D - \lambda_{-h})^2} + \frac{\lambda_h}{\mu'^2_h} - \frac{\lambda_h^2 \lambda_{-h}}{\mu'_B(\mu'_B - \lambda_{-h})^3} - \frac{\lambda_h^2 \lambda_{-h}}{\mu_D(\mu_D - \lambda_{-h})^3}. \tag{3.42}$$

In (3.42), $\lambda_{-h} = (N-1)\lambda_h$. *Proof:* See Appendix A.2. ■

Here, we denote γ_n as the ratio between $\Delta_{n,\text{gap}}$ and Ω_n, i.e., $\gamma_n = \frac{\Delta_{n,\text{gap}}}{\Omega_n}$, which indicates the error ratio if we use these bounds to approximate the average AoI of U_n. From Theorem 3.2, we find that γ_n is negligible, i.e., $\gamma_n \approx 0$, when the transmission rate and the computation rates of both the edge server and the local server at each UE are much larger than the packet generation rate. Under this condition, the average AoI is close to the average PAoI, i.e., $\Delta_n \approx \Omega_n$.

We now derive the optimal offloading ratio to minimize the average PAoI, which is approximated as the optimal offloading ratio to minimize the average AoI. The optimal offloading ratio, p_{opt}, to minimize the average PAoI should satisfy

$$p_{\text{opt}} = \arg\min_{p} \frac{p}{\mu_B - p\lambda} + \frac{1}{N}\sum_{n=1}^{N} \frac{1-p}{\mu_n - (1-p)\lambda_n} \tag{3.43}$$

$$\text{s.t. } \mu_B > p\lambda, \tag{3.43a}$$

$$\mu_n > (1-p)\lambda_n,\ n = 1,2,\cdots,N. \tag{3.43b}$$

Based on (3.43), we next analyze the optimal offloading ratio to minimize the average PAoI in the heterogeneous system.

Lemma 3.1. In the MEC system, the optimal offloading ratio, p_{opt}, to minimize the average PAoI is the solution to achieve

$$\frac{\mu_B}{(\mu_B - p_{\text{opt}}\lambda)^2} = \frac{1}{N}\sum_{n=1}^{N} \frac{\mu_n}{(\mu_n - (1-p_{\text{opt}})\lambda_n)^2}, \tag{3.44}$$

where p_{opt} satisfies $p_{\text{opt}} < \min\{\frac{\mu_B}{\lambda}, 1\}$ and $p_{\text{opt}} > \max\{\frac{\mu_n - \lambda_n}{\lambda_n}, 0\}$ for $n = 1,2,\cdots,N$. It is noted that (3.44) can be solved by using a number of numerical solvers, e.g., MATLAB.

Proof: See Appendix A.3. ∎

We note that it is indeed difficult, if not impossible, to obtain the closed-form expression for p_{opt} in non-homogeneous systems in Lemma 3.1. Nevertheless, a closed-form result is possible for homogeneous systems, where all UEs share the same packet generation rate λ_h and the same computation rate of the local server $\mu'_h = \frac{\mu_h}{1-p}$, i.e., $\lambda_n = \lambda_h$ and $\mu_n = \mu_h$, $\forall\, n$. Our analysis in the homogeneous case allows us to explicitly examine the impact of system parameters on the AoI performance[7]. In this homogeneous scenario, due to the complexity of the average AoI expression given in (3.24), it is hard to derive the closed-form expression for the optimal offloading ratio to minimize the average AoI. Thus, we resort to the average PAoI.

Theorem 3.3. In the homogeneous MEC system where all UEs share the same packet generation rate, i.e., $\lambda_n = \lambda_h$, and the same computation rate of the local server, i.e., $\mu_n = \mu_h$, $\forall\, n$,

[7] In addition, we note that the homogeneous scenario where all UEs share the same packet generation rate and computation rate may hold for some practical applications. For example, in vehicular networks, the packet generation rate for each vehicle is approximately the same and vehicles have very similar computation rate [112].

the optimal offloading ratio, p_{opt}, to minimize the average PAoI is derived as

$$p_{\text{opt}} = \begin{cases} 0, & \text{if } \mu_B \leq \dfrac{(\mu_h - \lambda_h)^2}{\mu_h}, \\ 1, & \text{if } \mu_h \leq \dfrac{(\mu_B - \lambda)^2}{\mu_B}, \\ \dfrac{\sqrt{\mu_B \mu_h} + \lambda_h - \mu_h}{\left(1 + N\sqrt{\frac{\mu_h}{\mu_B}}\right)\lambda_h}, & \text{otherwise.} \end{cases} \tag{3.45}$$

Proof: See Appendix A.4. ■

From Theorem 3.3, we see that the optimal offloading ratio increases as the computation rate of the edge server increases, but decreases as the number of UEs increases and the computation rate of the local server increases. It indicates that it is better to allocate more tasks to the edge server if the computation rate of the edge server is large and the number of UEs is small. However, if the number of UEs is large, UEs compete with each other in the computation queue of the edge server at the BS, which leads to the large waiting time of the packet in the computation queue of the edge server. In this case, it is better to allocate more tasks to the local server.

3.5 Numerical Results

In this section, we present numerical results to validate our analysis in Section 3.3 and 3.4. In particular, we present numerical results in the homogeneous case where all UEs share the same packet generation rate λ_h and the same local computation rate μ_h, i.e., $\lambda_n = \lambda_h$ and $\mu_n = \mu_h$, for $\forall n$. The simulations are conducted by the MATLAB and the system parameters are listed in the caption of each figure.

Fig. 3.3 plots the average AoI and PAoI of the MEC system versus the packet generation rate, λ_h. We first observe that the analytical average AoI and PAoI of the MEC system tightly match the simulation results, which demonstrates the correctness of our analytical result. We then observe that for all considered values of p, the average AoI and PAoI of the MEC system first decrease and then increases when λ_h increases, where the average AoI has a same trend with the average PAoI versus the packet generation rate, λ_h. This observation is due to the fact that the increase in λ_h has a two-fold effect on the average AoI and PAoI of the MEC system. When λ_h is small, this increase leads to a higher updating rate of packets, which decreases the average AoI and PAoI of the MEC system. When λ_h exceeds a certain threshold, its increase leads to the significant increase in the waiting time of a packet in computation queues and the transmission queue, thereby degrading the AoI performance. We further observe that for small λ_h, the average AoI and PAoI of the MEC system monotonically decreases with increasing p.

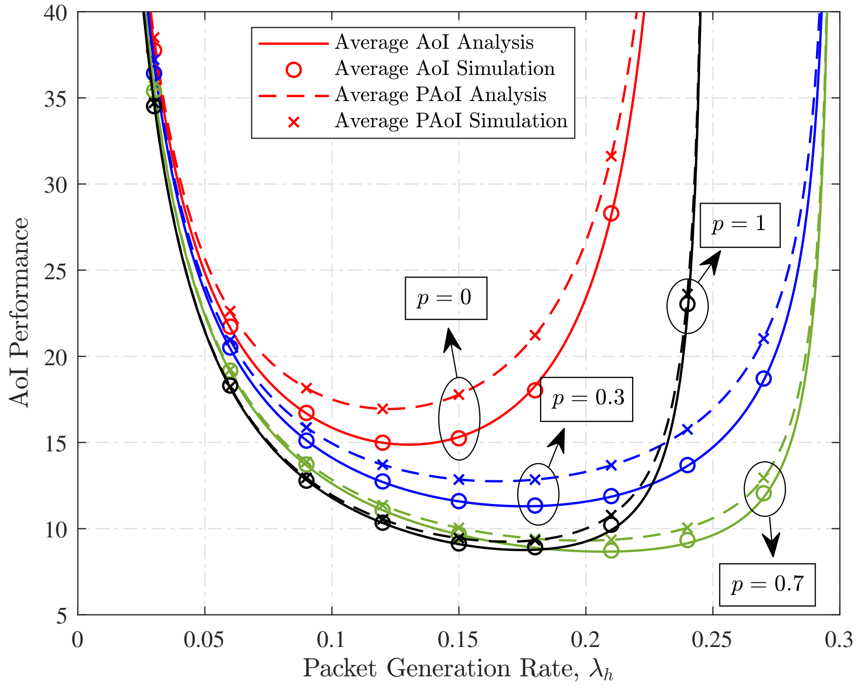


Figure 3.3: The average AoI and PAoI of the MEC system versus the packet generation rate, λ_h, with $N = 6$, $\mu_B = 1.5$, $\mu_h = 0.25$, and $\mu_D = 1.8$.

However, for large λ_h, the average AoI and PAoI of the MEC system firstly decreases and then increases with increasing p. This is because that when λ_h is small, the increase in p leads to a higher computation rate of the edge server, which reduces the average AoI and PAoI of the MEC system. When λ_h is very large, the increase in p results in a significantly long waiting time of a packet in the computation queue at the edge server, which increases the average AoI and PAoI of the MEC system. In addition, we find that the average AoI and PAoI is small in the partial computing scheme, but large in the other two schemes for large λ_h. This observation implies that by carefully partitioning the computational tasks, the partial computing scheme has a higher system stability than both the local computing and the edge computing schemes.

Fig. 3.4 plots the average AoI and PAoI of the MEC system versus the number of UEs, N. We first observe that when N increases, the average AoI and PAoI of the MEC system increase monotonically and this increase is faster for larger p. This is because that the increase in N results in the longer waiting time of a packet in both the transmission queue and the computation queue at the edge server, which increases the average AoI and PAoI of the MEC system. If p is large, the tasks computed by the edge server increases dramatically as N increases, which results in a long waiting time of a packet in the computation queue. We further observe that the edge computing scheme has a lower average AoI and PAoI than the local computing scheme

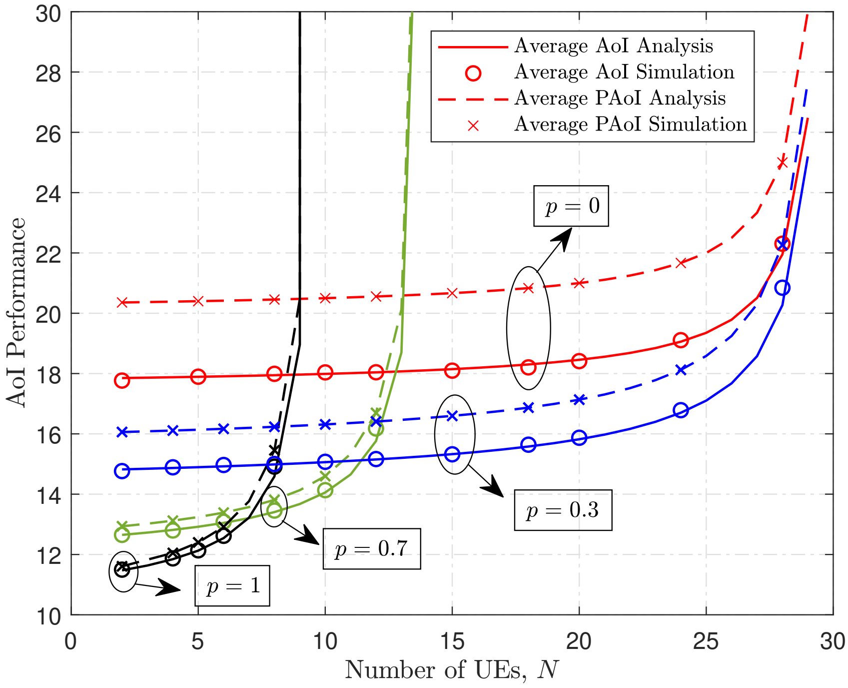


Figure 3.4: The average AoI and PAoI of the MEC system versus the number of UEs, N, with $\lambda_h = 0.1$, $\mu_B = 1$, $\mu_h = 0.2$, and $\mu_D = 3$.

and the partial computing scheme for small N, but a larger average AoI and PAoI for large N. This is because that for a small number of UEs, compared to local computing, edge computing can provide the higher computation rate via the powerful edge server, thereby decreasing the average AoI and PAoI of the MEC system. Differently, for a large number of UEs, allocating most computational tasks to the local server can avoid the long waiting time in the computation queue at the edge server, which is beneficial to the decrease in the average AoI and PAoI. This observation also indicates that the careful design of the offloading ratio in the partial computing scheme can effectively decrease the average AoI and PAoI, especially when the number of UEs is large.

Fig. 3.5 plots the average AoI of the MEC system versus the offloading ratio, p, for different μ_D and μ_B. We first observe that the average AoI first decreases and then increases as p increases. To understand the trend observed in these curves, we note that the increase in p leads to the decrease in the waiting time in the computation queue at the local server but the increase in the waiting time in the computation queue at the edge server. When p is small, its increase significantly reduces the waiting time of a packet at the local server, which decreases the average AoI of the MEC system. When the value of p increases up to a certain threshold, the long waiting time of a packet at the edge server dominates the average AoI, thereby result-

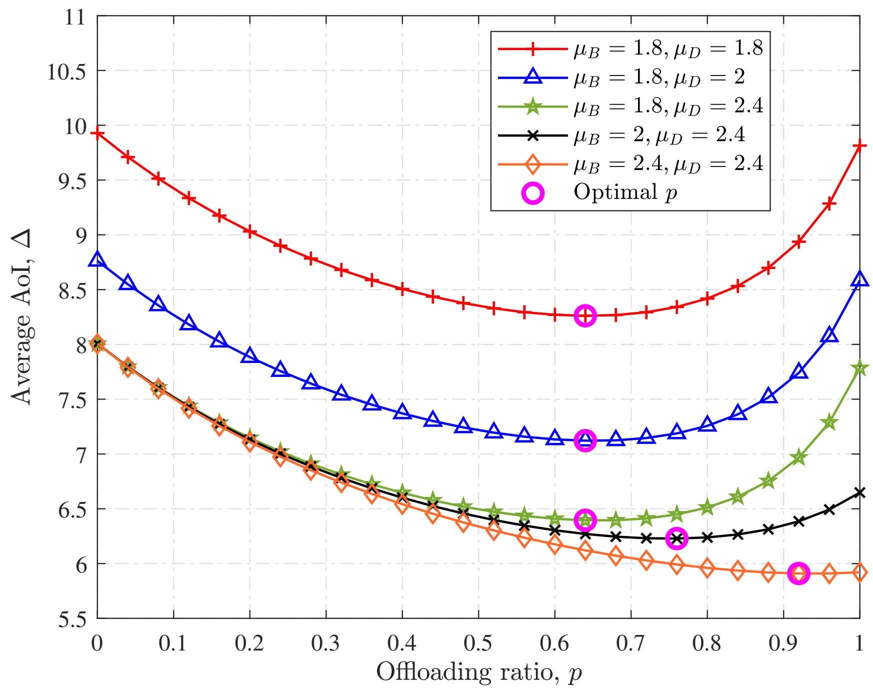


Figure 3.5: The average AoI of the MEC system versus the offloading ratio, p, with $N = 6$, $\lambda_h = 0.25$, and $\mu_h = 0.5$.

ing in the increase in the average AoI. We then observe that when μ_D increases, the average AoI of the MEC system decreases, while the optimal p almost remains the same value. This is because that as μ_D increases, the high transmission rate of the BS leads to reduced waiting time in the transmission queue, which decreases the average AoI. However, the transmission rate of the BS affects on the queuing delay of a packet in the transmission queue, while the optimal p depends on computation queues at both the edge server and the local server but independent of the transmission queue. We further observe that when μ_B increases, the average AoI of the MEC system decreases and the optimal p increases. This is because that the larger value of μ_B means the higher computation rate of the edge server, which can reduce the waiting time in the computation queue at the edge server and then lead to decreased average AoI. In addition, when the edge server has a higher computation rate, it can decrease the average AoI by allocating more computational tasks to the edge server.

Fig. 3.6 plots the upper bound and the lower bound on the average AoI in Theorem 3.2. We first observe that the average AoI is lower then the upper bound and higher then the lower bound given in Theorem 3.2, which demonstrates the correctness of our analytical result. We then observe that γ is small when λ_h is small and it is large for a large λ_h. This observation is due to the fact that the decrease in λ_h leads to the decrease in the gap between the upper

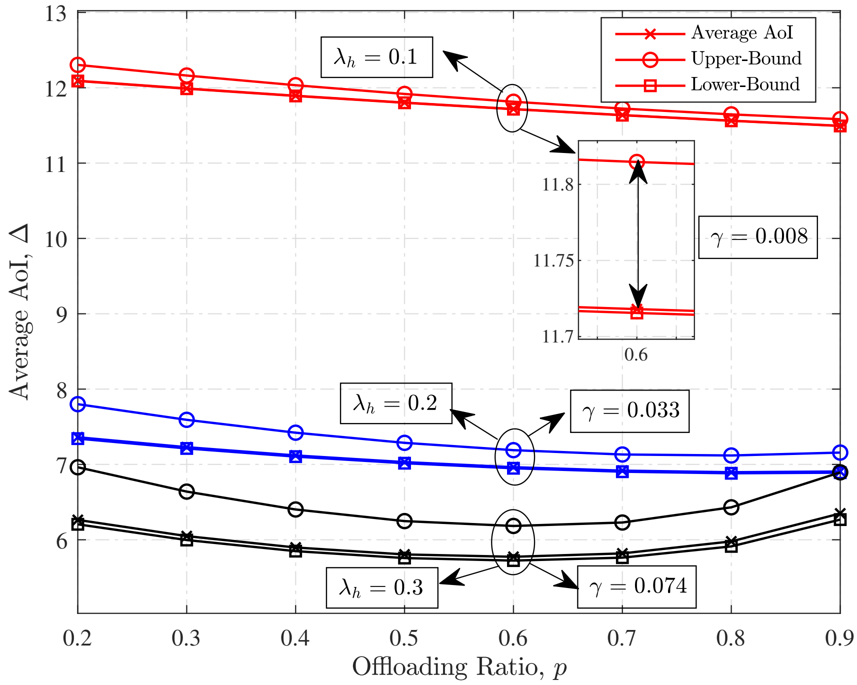


Figure 3.6: The average AoI of the MEC system versus the offloading ratio, p, with $N = 4$, $\mu_B = 1.5$, $\mu_h = 0.6$, and $\mu_D = 2$.

bound and the lower bound, which decreases γ. This observation implies that the expression for the average PAoI can be adopted as a reasonable approximation of the average AoI when the serving rate is much greater than the packet generation rate.

Fig. 3.7 plots the average AoI achieved by the partial computing scheme with the optimal offloading ratio, the partial computing scheme with the offloading ratio in Theorem 3.3, the local computing scheme, the edge computing scheme, and the partial computing scheme with $p = 0.5$ of the MEC system versus the packet generation rate, λ_h, for different N. We observe that the considered partial computing scheme with the offloading ratio in Theorem 3.3 achieves the lowest average AoI comparing with the local computing scheme, the edge computing scheme, and the partial computing scheme with $p = 0.5$. Additionally, we observe that the gap between the optimal average AoI and the average AoI with the offloading ratio in Theorem 3.3 is negligible no matter λ_h is small or large. It implies that the optimal offloading ratio, p_{opt}, to minimize the average PAoI can be adopted to minimize the average AoI of an MEC system, regardless of the packet generation rate.

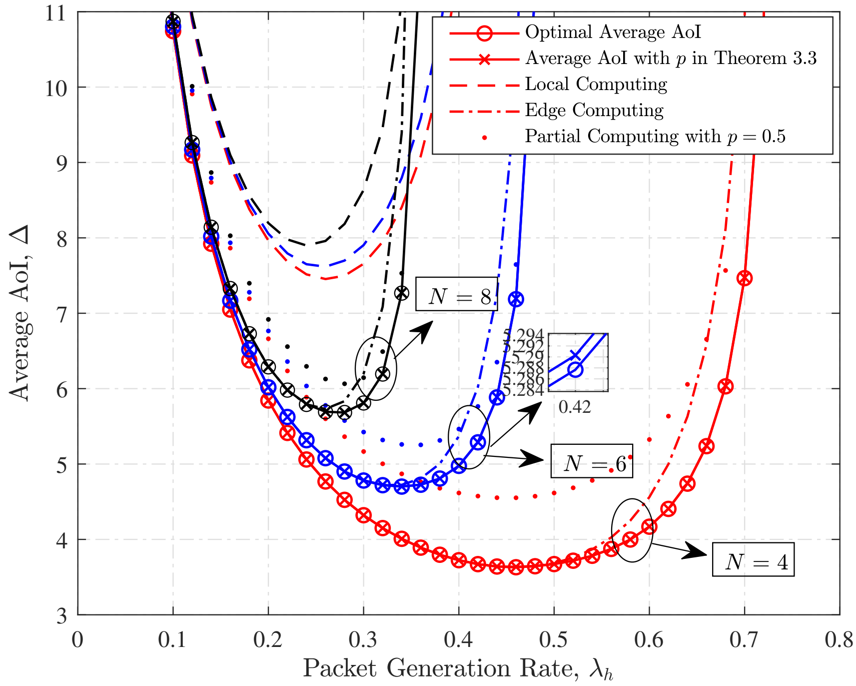


Figure 3.7: The average AoI of the MEC system versus the packet generation rate, λ_h, with $\mu_B = 2$, $\mu_h = 0.5$, and $\mu_D = 3$.

3.6 Summary

In this chapter, we analyzed the AoI performance of a multi-user MEC system. In this MEC system, we considered three computing schemes, i.e., the local computing scheme, the edge computing scheme, and the partial computing scheme. For these three computing schemes, we derived the closed-form expressions for the average AoI and the average PAoI, where packets are served in both the computation queues and the transmission queue according to the FCFS policy. Due to the complex expression for the average AoI, we further provided a simple expression for the upper bound and lower bound on the average AoI. Considering the sporadic packet generation pattern, we found that the gap between the upper bound and the lower bound is negligible, such that the average AoI can be approximated by the average PAoI. In addition, we derived the optimal offloading ratio to minimize the average PAoI and found that this offloading ratio can be adopted to minimize the average AoI.

Chapter 4

AoI Analysis of Short Packet Communications

4.1 Introduction

In this chapter, we investigate the AoI performance in point-to-point short packet communication systems. Although the existing studies [63, 64, 65, 66, 67, 68, 69, 70, 71, 72] have evaluated the impact of the packet blocklength and packet management strategies on the AoI performance in short packet communication systems, the impact of different encoding schemes on the AoI performance has not been evaluated, and the non-linear AoI performance of short packet communication systems remains untouched, which requires further analysis. In Section 4.2, to find an effective encoding scheme improving AoI performance in short packet communication systems, we analyze the impact of two encoding schemes, namely, the joint encoding scheme and the distributed encoding scheme, on the average AoI. We then analyze the average AoI penalty performance in short packet communication systems to gain deep insights into the joint requirements of system reliability and timeliness in Section 4.3.

4.2 The AoI of Short Packet Communications

In this section, we analytically assess the impact of two encoding schemes on the AoI performance of a short packet communication system. In this system, a transmitter transmits status updates received from multiple sensors to a receiver, where such status updates can be jointly encoded into a packet for transmission, referred to as the joint encoding scheme, or encoded into separate packets individually, referred to as the distributed encoding scheme. In these two encoding schemes, the packet generation follows the zero-waiting policy, where the transmitter generates a new packet immediately after finishing the transmission of the current packet. We derive new closed-form expressions for the average AoI achieved by both encoding schemes. Moreover, we theoretically determine the condition when one encoding scheme has a better AoI performance than the other. Using simulations, we demonstrate the accuracy of

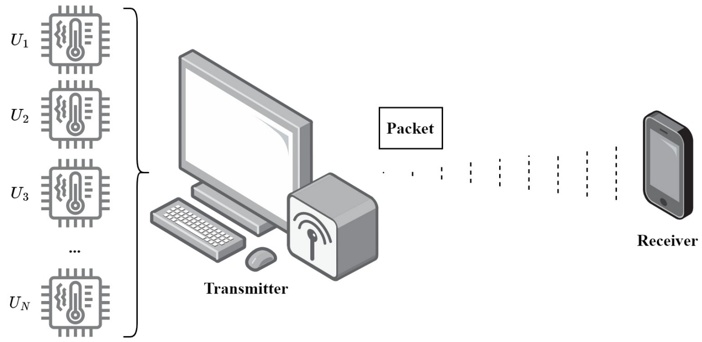


Figure 4.1: Illustration of our considered system where a transmitter collects status updates from N sensors and transmits the encoded packets to the receiver.

our analytical results and discuss the appropriate choice of the encoding scheme to decrease the average AoI based on the information redundancy and the number of sensors.

This section is organized as follows. In Section 4.2.1, the system model and the average AoI of the considered system are presented. The closed-form expressions for the average AoI are derived in Section 4.2.2. The numerical results and summary are provided in Section 4.2.3 and 4.2.4, respectively.

4.2.1 System Model and Average AoI

4.2.1.1 System Model

We consider a point-to-point wireless communication system as depicted in Fig. 4.1, where a transmitter transmits packets to a receiver. In this system, the transmitter collects status updates from N sensors and generates packets based on these collected status updates[1]. We denote the nth sensor by U_n, where $n = 1, 2, \cdots, N$. To ensure the freshness of packets, the zero-waiting policy is adopted, where a new packet is immediately generated once the transmitter finishes the transmission of the current packet. In this system, we introduce two encoding schemes, namely, the joint encoding scheme and the distributed encoding scheme.

In the joint encoding scheme, we assume that the joint status update from N sensors at the transmitter contains a fixed number of bits, denoted by L. Once the transmitter collects the status update, the transmitter encodes this L bits status update into a packet with the blocklength as M channel use (c.u.) and transmits this packet to the receiver. In the distributed encoding scheme, the transmitter collects the status update from sensors and follows the RR policy. In each time, the transmitter collects the fixed L_n bits status update from the sensor U_n and then encodes the L_n bits status update into a packet with M_n c.u.. Here, we assume that the N sensors are homogeneous sensors, where they have the same length of status updates and blocklength of packets, i.e., $L_n = L_h$ and $M_n = M_h$, $\forall n$. After the packet generation, the transmitter transmits the packet to the receiver. For the transmission, we define the coding rate, R, as the ratio

[1]We clarify that the status update collection process of N sensors is independent, but the collected data of N sensors may not be independent.

between the number of bits in the status update and the blocklength of the transmitted packet, i.e., $R = \frac{L}{M}$ in the joint encoding scheme and $R = \frac{L_h}{M_h}$ in the distributed encoding scheme.

We clarify that the information for the total N sensors are L bits in the joint encoding scheme and NL_h bits in the distributed encoding scheme. According to [113], the joint information is less than or equal to the sum of the individual information, i.e., $L \leq NL_h$. Thus, we denote the information redundancy over all the sensors by ϕ bits, i.e., $\phi = NL_h - L$.

As in [67], we assume an additive white Gaussian noise (AWGN) channel between the transmitter and the receiver. According to [62], the block error rate for the AWGN channel using finite block length coding can be approximated as

$$\varepsilon(l,m,\gamma) = Q\left(\frac{\frac{1}{2}\log_2(1+\gamma) - \frac{l}{m}}{\log_2(e)\sqrt{\frac{1}{2m}\left(1 - \frac{1}{(1+\gamma^2)}\right)}}\right), \tag{4.1}$$

where l is the number of bits in the status update, m is the blocklength of the transmitted packet, γ is the received SNR at the receiver, and $Q(x)$ is the Q-function. The expression for the block error rate in (4.1) is very tight for short packet communications when $m \geq 100$ [62] and is close to 0 for a significantly large m. For the packet transmission in our considered system, we define the transmission time for each c.u. as the unit time for analytical simplicity.

4.2.1.2 AoI Formulation

In this subsection, we formulate the expression for the average AoI of the considered time-slotted system where T_u is treated as the unit time.

We denote $\Delta_n(t)$ as the AoI of U_n at time slot t. We assume that the observation begins at $t = 0$ with the AoI of $\Delta_n(0)$. Then, we express the AoI at time t as

$$\Delta_n(t) = t - u_n(t), \tag{4.2}$$

where $u_n(t)$ is the generation time of the most recently received status update at receiver from U_n at time t. Based on the zero-waiting policy, we note that the generation time of the most recently received packet is the time that the transmitter finishes its transmission. We then denote $\Delta(t)$ as the overall AoI of the system at time t. Based on (4.2), $\Delta(t)$ is defined as

$$\Delta(t) = \max_n \{\Delta_n(t)\}, \tag{4.3}$$

which is the maximum AoI of all the sensors at time t. Then, the average AoI over the observation window of T time slots is calculated as

$$\Delta = \frac{1}{T}\sum_{t=0}^{T-1}\Delta(t). \tag{4.4}$$

We next discuss the AoI expressions for the joint encoding scheme and the distributed encoding scheme separately, due to the fundamental differences in these two schemes.

1. In the joint encoding scheme, since the transmitter collects the joint status update from all the sensors, the AoI of different sensors at any time is the same, i.e.,

$$\Delta(t) = \Delta_{n_1}(t) = \Delta_{n_2}(t), \tag{4.5}$$

where n_1 and $n_2 \in \{1,2,\cdots,N\}$. For the convenience of the following analysis, we consider that T time slots are grouped into K frames and each frame has M time slots, i.e., $T = KM$. By averaging $\Delta(t)$ over a sufficiently long observation window with $T \to \infty$, we obtain the average AoI as

$$\Delta_J = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\Delta(t) = \lim_{K\to\infty}\frac{1}{M}\sum_{m=0}^{M-1}\frac{1}{K}\sum_{k=0}^{K-1}\Delta(kM+m). \tag{4.6}$$

Since $\Delta(kM+m) = \Delta(kM)+m$, we simplify (4.6) as

$$\Delta_J = \lim_{K\to\infty}\frac{1}{M}\sum_{m=0}^{M-1}\frac{1}{K}\sum_{k=0}^{K-1}(\Delta(kM)+m) = \mathrm{AoI}_0 + \frac{1}{M}\sum_{m=0}^{M-1}m = \mathrm{AoI}_0 + \frac{M-1}{2}, \tag{4.7}$$

where $\mathrm{AoI}_0 = \lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}\Delta(kM)$.

2. In the distributed encoding scheme, we consider T time slots grouped into K frames and each frame has NM_h time slots, i.e., $T = KNM_h$. Then, the average AoI is calculated as

$$\begin{aligned}\Delta_D &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\Delta(t) = \lim_{K\to\infty}\frac{1}{NM_h}\sum_{m=0}^{NM_h-1}\frac{1}{K}\sum_{k=0}^{K-1}\Delta(kNM_h+m)\\ &= \lim_{K\to\infty}\frac{1}{M_h}\sum_{m=0}^{M_h-1}\frac{1}{N}\sum_{n=1}^{N}\frac{1}{K}\sum_{k=0}^{K-1}\Delta((kN+n)M_h+m) = \frac{1}{M_h}\sum_{m=0}^{M_h-1}\frac{1}{N}\sum_{n=1}^{N}\mathrm{AoI}_{n,m},\end{aligned} \tag{4.8}$$

where $\mathrm{AoI}_{n,m} = \lim_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}\Delta((kN+n)M_h+m)$. Due to the same length of status updates and blocklength of packets for N sensors, we note that

$$\mathrm{AoI}_{n_1,m} = \mathrm{AoI}_{n_2,m}, \tag{4.9}$$

for any $n_1, n_2 \in \{1,2,\cdots,N\}$. Combining (4.9) with $\Delta(knM_h+m) = \Delta(knM_h)+m$, we simplify (4.8) as

$$\Delta_D = \frac{1}{M_h} \sum_{m=0}^{M_h-1} \mathrm{AoI}_{n,m} = \mathrm{AoI}_{N,0} + \frac{M_h - 1}{2}. \tag{4.10}$$

4.2.2 Closed-form Analysis of Average AoI

In this section, we derive the closed-form expressions for the average AoI achieved by both encoding schemes. We first focus on the joint encoding scheme and present its average AoI in the following theorem.

Theorem 4.1. In the joint encoding scheme with the received SNR, γ, the closed-form expression for the average AoI is derived as

$$\Delta_J = \frac{M}{1-\varepsilon_J} + \frac{M-1}{2}, \tag{4.11}$$

where $\varepsilon_J = \varepsilon(L, M, \gamma)$ and $\varepsilon(\cdot,\cdot,\cdot)$ is given in (4.1).

Proof: In this scheme, based on (4.7), we need to derive AoI_0 to obtain the average AoI. Note that the AoI expression is periodic with the period M, which is given by

$$\Delta\left((k+1)M\right) = \begin{cases} M, & \mathcal{C}_k, \\ \Delta(kM) + M, & \text{otherwise}, \end{cases} \tag{4.12}$$

where $\mathcal{C}_k$ is the event that the packet generated at $t = kM$ is successfully received by the receiver. By exploiting the probability of the occurrence of $\mathcal{C}_k$, denoted by $\Pr(\mathcal{C}_k)$, we obtain

$$\mathbb{E}[\Delta((k+1)kM)] = \Pr(\mathcal{C}_k)M + (1-\Pr(\mathcal{C}_k))\mathbb{E}[\Delta(kM)+M], \tag{4.13}$$

where $\Pr(\mathcal{C}_k) = 1 - \varepsilon_J$. According to [18], we obtain

$$\mathrm{AoI}_0 = \mathbb{E}[\Delta(kM)] = \mathbb{E}[\Delta((k+1)M)]. \tag{4.14}$$

Combining (4.14) with (4.13), we derive AoI_0 as

$$\mathrm{AoI}_0 = \frac{M}{1-\varepsilon_J}. \tag{4.15}$$

By substituting (4.15) into (4.6), we obtain (4.11). ■

We then derive and present the closed-form expression for the average AoI achieved by the distributed encoding scheme in the following theorem.

Theorem 4.2. In the distributed transmission scheme with the received SNR, γ, the closed-form expression for the average AoI is derived as

$$\Delta = \sigma N M_h + \psi M_h + \frac{M_h - 1}{2}, \tag{4.16}$$

where

$$\sigma = \sum_{n=0}^{N} \binom{N}{n} \frac{(-\varepsilon_D)^n}{1-\varepsilon_D^n} \tag{4.17}$$

and

$$\psi = (1-\varepsilon_D) \sum_{n=1}^{N} (N-n+1) \sum_{n_1=0}^{n-1} \sum_{n_2=0}^{N-n} \binom{n-1}{n_1} \binom{N-n}{n_2} \frac{(-1)^{n_1+n_2} \varepsilon_D^{n_2}}{1-\varepsilon_D^{n_1+n_2+1}}, \tag{4.18}$$

with $\varepsilon_D = \varepsilon(L_h, M_h, \gamma)$ and $\varepsilon(\cdot,\cdot,\cdot)$ is given in (4.34).

Proof: In this scheme, based on (4.10), we need to derive $\mathrm{AoI}_{N,0}$ to obtain the average AoI. Note that the AoI expression is periodic with period NM_h and the AoI of U_n at $t = kNM_h$ is given by

$$\Delta_n(kNM_h) = \delta_n N M_h + (N-n+1) M_h, \tag{4.19}$$

where δ_n is the number of consecutive transmission failures for packets from U_n. We find that δ_n follows a geometric distribution, whose PMF is given by

$$\Pr(\delta_n = \delta) = \varepsilon_D^{\delta}(1-\varepsilon_D). \tag{4.20}$$

Based on [18], we obtain

$$\mathbb{E}[\Delta_n(kNM_h)] = \mathbb{E}[\Delta_n((k+1)NM_h)]. \tag{4.21}$$

According to (4.3), (4.9), and (4.19), we obtain $\mathrm{AoI}_{N,0}$ in (4.10) as

$$\begin{aligned} \mathrm{AoI}_{N,0} &= \max_{n} \mathbb{E}[\Delta_n(kNM_h)] = \max_{\delta_n} \mathbb{E}\left[\delta_n N M_h + (N-n+1) M_h\right] \\ &= \mathbb{E}[\delta_{max} N M_h] + \mathbb{E}\left[(N-n^*+1) M_h\right], \end{aligned} \tag{4.22}$$

where $\delta_{max} = \max\{\delta_n\}$ and $n^* = \min\left\{\arg\max_{n} \delta_{max}\right\}$. It shows that the system AoI is dominated by the largest number of consecutive transmission failures in this scheme. Accordingly, the cumulative distribution function (CDF) of δ_{max} is given as

$$F_{\delta_{max}}(\delta) = \Pr(\delta_{max} < \delta) = (1-\varepsilon_D^{\delta})^N. \tag{4.23}$$

Based on the CDF of δ_{max}, we obtain the PMF of δ_{max} as

$$\Pr(\delta_{max}=\delta) = \left(1-\varepsilon_D^{\delta+1}\right)^N - \left(1-\varepsilon_D^{\delta}\right)^N. \tag{4.24}$$

According to the PMF of δ_{max} given in (4.24), the first item in (4.22) is calculated as

$$\mathbb{E}[\delta_{max}NM_h] = \sum_{\delta=0}^{\infty} \delta NM_h \Pr(\delta_{max}=\delta) = NM_h \sum_{\delta=0}^{\infty} \delta \left(\left(1-\varepsilon_D^{\delta+1}\right)^N - \left(1-\varepsilon_D^{\delta}\right)^N\right). \tag{4.25}$$

We note that

$$\sum_{\delta=0}^{\infty} \delta \left(\left(1-\varepsilon_D^{\delta+1}\right)^N - \left(1-\varepsilon_D^{\delta}\right)^N\right) = \sum_{\delta=0}^{\infty} \delta \sum_{n=0}^{N} \binom{N}{n} \left(\left(-\varepsilon_D^{\delta+1}\right)^n - \left(-\varepsilon_D^{\delta}\right)^n\right) = \sigma. \tag{4.26}$$

Hence, we obtain

$$\mathbb{E}[\delta_{max}NM_h] = \sigma NM_h. \tag{4.27}$$

We then calculate the second item in (4.22). We first note that the number of consecutive transmission failures of the packets from U_{n^*} is the largest number of consecutive transmission failures of the packets from all the sensors, i.e., $\delta_{n^*} = \delta_{max}$. We also note that the number of consecutive transmission failures of the packets from U_{n^*} is greater than the number of consecutive transmission failures of the packets from U_k, $\forall k < n^*$, i.e., $\delta_{n^*} > \delta_k$. Thus, we obtain

$$\Pr(n^*=n) = \sum_{\delta=0}^{\infty} \prod_{k=1}^{n-1} \Pr(\delta_k < \delta) \prod_{k=n+1}^{N} \Pr(\delta_k \le \delta) \Pr(\delta_n = \delta). \tag{4.28}$$

According to $\Pr(\delta_k < \delta) = 1-\varepsilon_D^{\delta}$, $\Pr(\delta_k \le \delta) = 1-\varepsilon_D^{\delta+1}$, and (4.20), we obtain

$$\Pr(n^*=n) = \sum_{\delta=0}^{\infty} \left(1-\varepsilon_D^{\delta}\right)^{n-1} \left(1-\varepsilon_D^{\delta+1}\right)^{N-n} \varepsilon_D^{\delta}(1-\varepsilon_D). \tag{4.29}$$

Thus, the second item in (4.22) is calculated as

$$\mathbb{E}\left[(N-n^*+1)M_h\right] = \sum_{n=1}^{N} \Pr(n^*=n)(N-n+1)M_h = \psi M_h. \tag{4.30}$$

By substituting (4.27) and (4.30) into (4.22), we obtain the expression for $\mathrm{AoI}_{N,0}$ as given in (4.22). In addition, by substituting the expression for $\mathrm{AoI}_{N,0}$ into (4.10), we obtain the final result for the distributed encoding scheme. ■

Based on Theorem 4.1 and Theorem 4.2, we obtain the expressions for the average AoI achieved by the joint encoding scheme and the distributed encoding scheme, respectively.

Since the block error rate is the function of the blocklength of a packet, there would exist the optimal blocklength of a packet to minimize their average AoI for both encoding schemes. Due to the complexity of the block error rate given in (4.1), it is hard to derive closed-form expressions for the optimal blocklength for the proposed encoding schemes. Thus, we resort to numerical methods to find the optimal blocklength, which are the solutions to the equation $\frac{\mathrm{d}\Delta_J}{\mathrm{d}M} = 0$ in the joint encoding scheme and the equation $\frac{\mathrm{d}\Delta_D}{\mathrm{d}M_h} = 0$ in the distributed encoding scheme.

We then compare the average AoI performance achieved by both encoding schemes by considering the information redundancy ϕ. Here, we consider the same coding rate, R, in the joint encoding scheme and the distributed encoding scheme, i.e., $R = \frac{L}{M} = \frac{L_h}{M_h}$. We assume that the coding rate leads to a low block error rate in both encoding schemes, which is due to the fact that high reliable communication technologies are typically adopted in practice. Theorem 4.3 compares the average AoI between two encoding schemes based on the information redundancy.

Theorem 4.3. In the system under the assumption of the same coding rate, R, and low block error rate for both encoding schemes, the joint encoding scheme has a better AoI performance than the distributed encoding scheme, if $\phi \geq \phi_0$, and the distributed encoding scheme has a better AoI performance than the joint encoding scheme, if $\phi < \phi_0$. Specifically, ϕ_0 is obtained as

$$\phi_0 = \frac{(3-2\sigma)N - 2\psi - 1}{3} L_h. \tag{4.31}$$

Proof: According to [114], the block error rate in the joint encoding scheme can be approximated as $\varepsilon_J \approx \varepsilon_D^N$. Once ε_D is low, the block error rate ε_J in the joint encoding scheme is negligible. Hence, the average AoI difference between the joint encoding scheme and the distributed encoding scheme is approximated as

$$\Delta_{\text{Diff}} = \Delta_J - \Delta_D \approx \left(\frac{3}{2}\left(N - \frac{\phi}{L_h}\right) - \left(N\sigma + \frac{1}{2} + \psi\right)\right) M_h. \tag{4.32}$$

Based on (4.32), we obtain $\Delta_{\text{Diff}} \leq 0$ when $\phi \geq \phi_0$, and $\Delta_{\text{Diff}} > 0$ when $\phi < \phi_0$.

■

Theorem 4.3 reveals that the joint encoding scheme is better when we can compress much information into a joint packet, but the distributed encoding scheme is better when there is low correlation among the information from different sensors. Moreover, comparing with the distributed encoding scheme, the joint encoding scheme requires an extra data processing process to eliminate information redundancy. While this step enhances data efficiency, it also introduces computational overhead, increasing the average AoI. We note that the computation

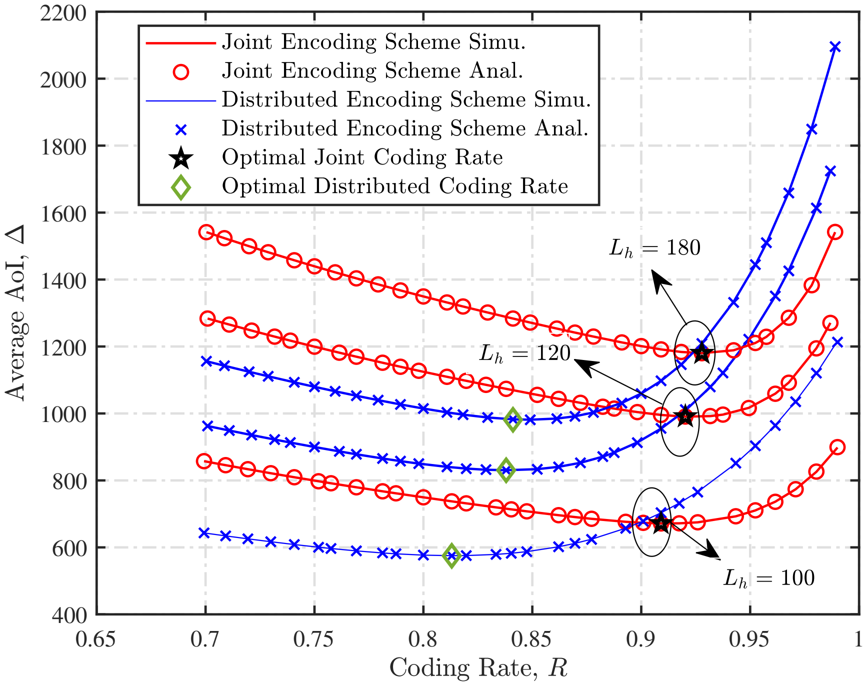


Figure 4.2: The average AoI of the considered system versus the coding rate, R, with $N = 4$, $\gamma = 3$, and $\phi = 0$.

time of the data processing process has a linear impact on the average AoI, which is calculated as

$$\Delta_{J,\text{Process}} = \Delta_J + T_{\text{Process}}, \tag{4.33}$$

where T_{Process} is the computation time of data processing process and it is inversely proportional to the computation rate. The limited computational resource leads to a large T_{Process}, degrading the average AoI of the joint encoding scheme. It reveals that the joint encoding scheme should be selected when the transmitter has a large amount of computational resources.

4.2.3 Numerical Results

In this section, we present numerical results and evaluate the impact of various parameters, including the coding rate, the number of bits in a status update, the number of sensors, and the information redundancy on the average AoI achieved by both encoding schemes in our considered system.

Fig. 4.2 plots the average AoI of the considered system versus the coding rate, R. We first observe that the analytical average AoI precisely matches the simulation results, which

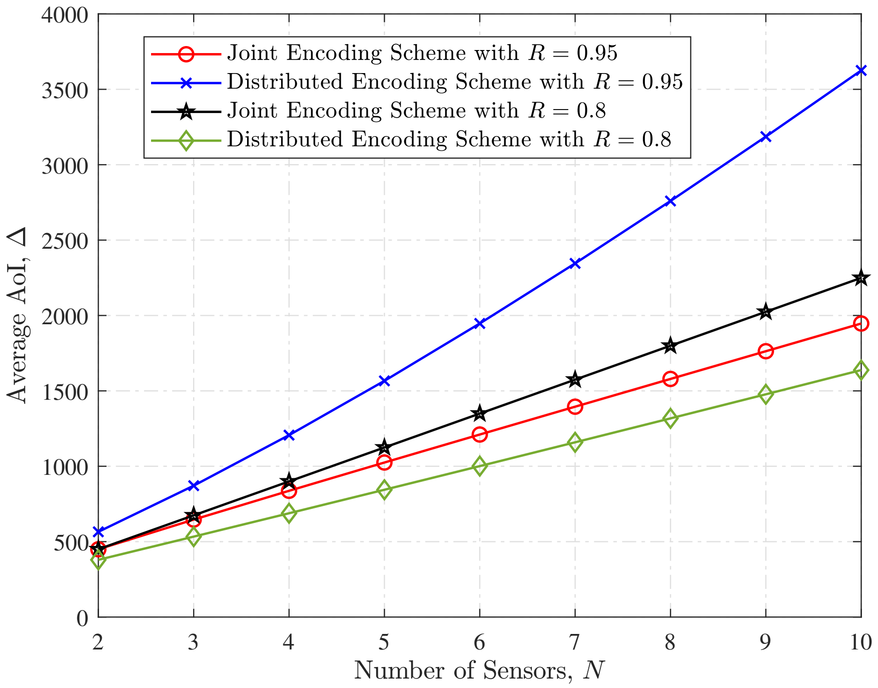


Figure 4.3: The average AoI of the considered system versus the number of sensors, N, with $L_h = 120$, $\gamma = 3$, and $\phi = 0$.

demonstrates the correctness of our analytical results in Theorem 4.1 and Theorem 4.2. We then observe that for both encoding schemes, the average AoI first decreases and then increases when R increases. This observation is due to the fact that the increase in R has a two-fold effect on the average AoI via the blocklength of a packet and the block error rate. When R is small, its increase leads to a smaller blocklength of packets, which decreases the average AoI of the considered system. When R exceeds a certain threshold, its increase leads to the significant increase in the block error rate, thereby degrading the AoI performance. We further observe that the optimal coding rate, which minimizes the average AoI, increases when the number of bits in a status update increases in both encoding schemes. This is because that the increase in the number of bits in a status update leads to the increase in the transmission time and the decrease in the block error rate. When the number of bits in a status update is low, the decrease in the block error rate dominantly and positively affects the average AoI. When the number of bits in a status update is high, the increase in the transmission time dominants the average AoI, thereby degrading the AoI performance.

Fig. 4.3 plots the average AoI of the considered system versus the number of sensors, N. We first observe that when N increases, the average AoI achieved by both encoding schemes increases monotonically. This is because that the increase in N results in the longer trans-

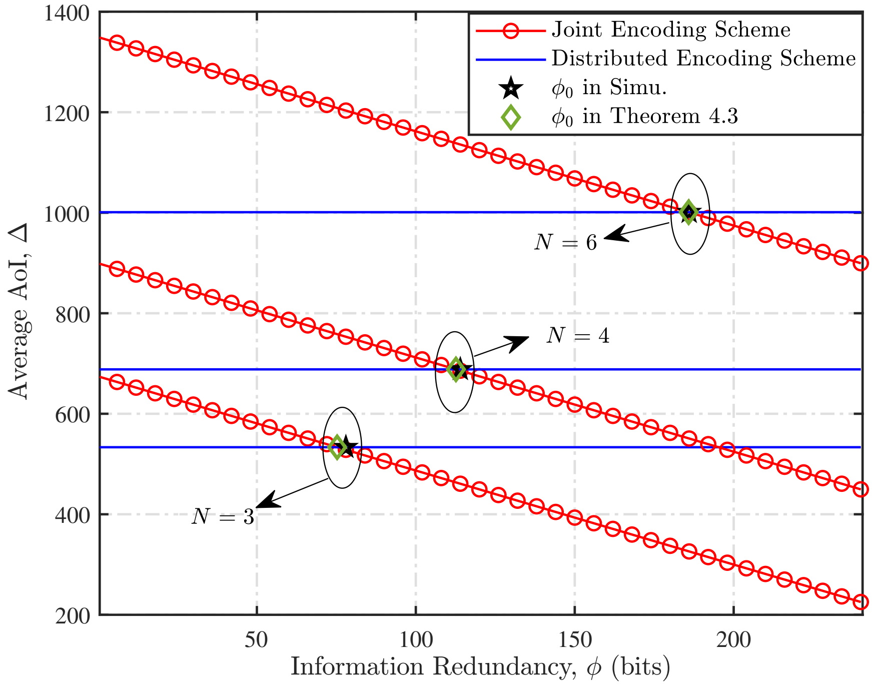


Figure 4.4: The average AoI of the considered system versus the information redundancy, ϕ, with $L_h = 120$, $\gamma = 3$, and $R = 0.8$.

mission time of status updates from all the sensors, which increases the average AoI of the system. We then observe that the joint encoding scheme achieves a lower average AoI than the distributed encoding scheme for a high coding rate, but a larger average AoI for a low coding rate. This is because that the block error rate increases when the coding rate increases. This block error rate has a more pronounced impact on the average AoI for the distributed encoding scheme than the joint encoding scheme.

Fig. 4.8 plots the average AoI of the considered system versus the information redundancy, ϕ. We first observe that the analytical ϕ_0 tightly matches the simulation result, which demonstrates the correctness of our analytical result in Theorem 4.3. It indicates that we can adopt Theorem 4.3 to find the optimal encoding scheme to minimize the average AoI based on the information redundancy. We then observe that the average AoI decreases when ϕ increases for the joint encoding scheme for different N. This is because that the increase in ϕ leads to the small number of bits in status updates for the transmission in the joint encoding scheme. Hence, the blocklength of a packet for the transmission decreases, which decreases the average AoI of the system. We further observe that ϕ_0 increases as the number of sensors increases. This observation is due to the fact that when N is large, the distributed encoding scheme significantly increases the update frequency of the states received at the receiver, compared with the joint encoding scheme, which decreases the average AoI. This observation also implies that it is better to select the joint encoding scheme when the number of sensors is small, but select the distributed encoding scheme when the number of sensors is large.

4.2.4 Summary

This section considered a point-to-point system where a transmitter generates packets and transmits them to a receiver under an unreliable channel. We analyzed the average AoI achieved by two encoding schemes, i.e., the joint encoding scheme and the distributed encoding scheme, under a finite packet block length model. By using simulation results, we demonstrated the accuracy of our analysis and showed that the block error rate has a more significant impact on the average AoI for the distributed encoding scheme than the joint encoding scheme. In addition, we found that the distributed encoding scheme has a better AoI performance than the joint encoding scheme in a highly reliable communication system with a large number of sensors.

4.3 Average AoI Penalty of Short Packet Communications with Packet Management

In this section, we analytically assess the impact of three packet management strategies on the non-linear AoI performance of a point-to-point short packet communication system. In this system, the transmitter generates status updates and transmits them to the receiver via an unreliable channel. We examine three packet management strategies, namely, the NPNB strategy, the NPOB strategy, and the preemption strategy. To characterize the level of the receiver's dissatisfaction on outdated status update, we derive new closed-form expressions for the average AoI penalty achieved by three packet management strategies, where a generalized α-β AoI penalty function [115] is adopted. Aided by simulations, we demonstrate the accuracy of our analytical results and find the relationship between the requirement of the transmission reliability and the toleration on the outdated date.

This section is organized as follows. In Section 4.3.1, the system model and the average AoI penalty of the considered system are presented. The closed-form expressions for the average AoI penalty are derived in Section 4.3.2. The numerical results and summary are provided in Section 4.3.3 and 4.3.4, respectively.

4.3.1 System Model and Average AoI Penalty

4.3.1.1 System Description

We consider a point-to-point wireless communication system, where a transmitter transmits packets to a receiver. In this system, the transmitter generates status updates according to a

Poisson process with the rate of λ. Then the transmitter encodes the generated status updates into packets and transmits the encoded packets to the receiver. We assume that each status update contains a fixed number of bits, denoted by L. Once the transmitter generates a status update, it encodes this L-bits information into a packet with the blocklength of M c.u. and transmits this packet to the receiver. For this transmission, we define the coding rate, R, as the ratio between the number of bits in the status update and the blocklength of the transmitted packet, i.e., $R = L/M$.

To ensure the freshness of packets, we assume that the LCFS queuing policy is adopted and there is no feedback signal sent by the receiver. When a new status update is generated and the transmitter is in the idle state, it encodes the status update into a packet and transmits this packet to the receiver. In this system, we examine three packet management strategies, detailed as follows:

1. NPNB strategy: In this strategy, the buffer size is 0. If a new status update is generated at the transmitter whilst the transmitter is transmitting a packet, the transmitter discards the new status update and keeps the current transmission. After transmitting the current packet, the transmitter waits for a new status update for transmission.

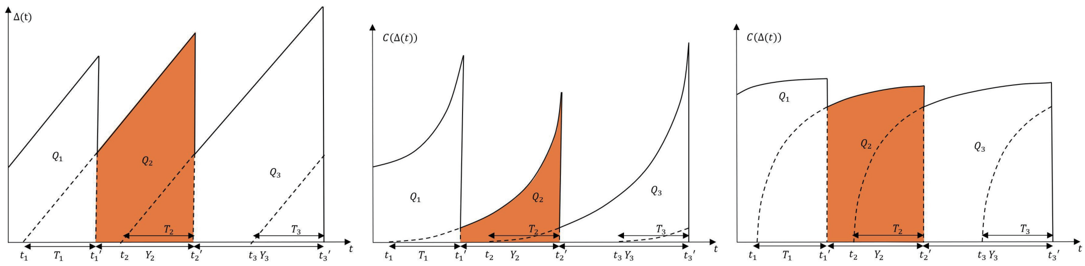
2. NPOB strategy: In this strategy, the buffer size is 1. If a new status update is generated at the transmitter whilst the transmitter is transmitting a packet, the transmitter stores the newly generated status update in the buffer and keeps the current transmission. After transmitting the current packet, the transmitter immediately removes the status update from the buffer and transmits this status update to the receiver.

3. Preemption strategy: If a new status update is generated at the transmitter whilst the transmitter is transmitting a packet, the transmitter discards the current transmission and starts to transmit the newly generated status update.

We assume an AWGN channel between the transmitter and the receiver, which is a standard assumption in the literature, e.g., [67]. According to [62], the block error rate for the AWGN channel using finite block length coding can be approximated as

$$\varepsilon(l,m,\gamma) = Q\left(\frac{\frac{1}{2}\log_2(1+\gamma) - \frac{l}{m}}{\log_2(e)\sqrt{\frac{1}{2m}\left(1-\frac{1}{(1+\gamma^2)}\right)}}\right), \tag{4.34}$$

where l is the number of bits in the status update, m is the blocklength of the transmitted packet, γ is the received SNR at the receiver, and $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\mathrm{d}t$ is the Q-function. The block error rate expression in (4.34) is very tight for short packet communications when $m \geq 100$ [62]. For the packet transmission in our considered system, we define the transmission time for each c.u. as the unit time for analytical simplicity.



(a) (b) (c)

Figure 4.5: Illustration of three AoI penalty functions with (a) linear AoI, (b) $\alpha > 0$ and $\beta > 0$, and (c) $\alpha < 0$ and $\beta < 0$.

4.3.1.2 Formulation of AoI Penalty

In this subsection, we formulate the expression for the average AoI penalty of the considered system. To this end, we first denote $\Delta(t)$ as the AoI at time t. Fig. 4.5(a) plots a sample variation $\Delta(t)$ as a function of t. We assume that the observation begins at $t = 0$ with the AoI of $\Delta(0)$. From Fig. 4.5(a), we express the AoI at time t as

$$\Delta(t) = t - u(t), \tag{4.35}$$

where $u(t)$ is the generation time of the most recently received status update at the receiver at time t.

Apart from the linear AoI, we also adopt the α-β AoI penalty function [115] in this work to characterize the receiver's dissatisfaction level on outdated data. Mathematically, this function is given by

$$C(\Delta(t)) = \beta\left(e^{\alpha\Delta(t)} - 1\right). \tag{4.36}$$

We find that $C(\Delta(t))$ is an exponentially shaped function when $\alpha > 0$ and $\beta > 0$, as shown in Fig. 4.5(b), or a logarithmically shaped function when $\alpha < 0$ and $\beta < 0$, as shown in Fig. 4.5(c) [115]. We note that the α-β AoI penalty function is a generalized function as the linear AoI is a special case of $C(\Delta(t))$ with $\lim \alpha \to 0$ and $\beta = 1/\alpha$.

We denote P_j as the jth successfully received packet generated after time $t = 0$, $j = 1, 2, \cdots$. We then denote ζ_j as the time interval from the received time of P_{j-1} to the received time of P_j and denote T_j as the time interval from the generation time of P_j to the received time of P_j. Thus, we obtain

$$\zeta_j = t'_j - t'_{j-1} \text{ and } T_j = t'_j - t_j, \tag{4.37}$$

where t_j is the generation time of P_j and t'_j is the time that P_j is received.

According to [115], the average AoI penalty is given by

$$\overline{C}=\frac{\mathbb{E}[Q_j]}{\mathbb{E}[\zeta_j]}, \tag{4.38}$$

where Q_j is the area under curve in Fig. 4.5. In (4.38), the area Q_j is calculated as

$$\begin{aligned} Q_j &= \int_{t'_{j-1}}^{t'_j} C(\Delta(t))\mathrm{d}t = \int_0^{\zeta_j+T_{j-1}} C(\Delta(t))\mathrm{d}t - \int_0^{T_{j-1}} C(\Delta(t))\mathrm{d}t \\ &= \frac{\beta}{\alpha} e^{\alpha T_{j-1}}\left(e^{\alpha\zeta_j}-1\right)-\beta\zeta_j. \end{aligned} \tag{4.39}$$

Since the LCFS queuing policy is adopted, ζ_j is independent of T_{j-1}. By substituting (4.39) into (4.38), and using the fact that ζ_j is independent of T_{j-1}, we obtain

$$\overline{C}=\frac{\beta\mathbb{E}\left[e^{\alpha T_{j-1}}\right]\left(\mathbb{E}\left[e^{\alpha\zeta_j}\right]-1\right)}{\alpha\mathbb{E}\left[\zeta_j\right]}-\beta. \tag{4.40}$$

We next analyze $\overline{C}$ achieved by three strategies.

4.3.2 Analysis of Average AoI Penalty

In this section, we derive the closed-form expressions for the average AoI penalty achieved by three packet management strategies. We first focus on the NPNB strategy and present its average AoI penalty in the following theorem.

Theorem 4.4. Considering the NPNB strategy with the received SNR, γ, the closed-form expression for the average AoI penalty is derived as

$$\overline{C}_{\mathrm{NPNB}}=\frac{\beta e^{\alpha M}\left(\lambda^2\left(1-\varepsilon_\gamma\right)\left(e^{\alpha M}-1\right)+\alpha\right)}{\alpha\left(\lambda M+1\right)\left(\lambda\left(1-\varepsilon_\gamma e^{\alpha M}\right)-\alpha\right)}-\beta, \tag{4.41}$$

where $\varepsilon_\gamma=\varepsilon(L,M,\gamma)$ and $\varepsilon(\cdot,\cdot,\cdot)$ is given in (4.34). *Proof:* Based on (4.40), we need to derive $\mathbb{E}[e^{\alpha T_{j-1}}]$, $\mathbb{E}\left[e^{\alpha\zeta_j}\right]$, and $\mathbb{E}[\zeta_j]$ to obtain the average AoI penalty. Since $T_{j-1}=M$ for the NPNB strategy, we first obtain

$$\mathbb{E}\left[e^{\alpha T_{j-1}}\right]=e^{\alpha M}. \tag{4.42}$$

We then derive $\mathbb{E}\left[e^{\alpha\zeta_j}\right]$ and $\mathbb{E}[\zeta_j]$. Here, we denote H_j as the number of packets transmitted to the receiver after P_{j-1} is received but by P_j is received, and $P_{j,h}$ as the hth packet transmitted to the receiver after P_{j-1} is received, where $P_{j,H_j} = P_j$. We note that

$$\zeta_j = \sum_{h=1}^{H_j} B_{j,h}, \tag{4.43}$$

where $B_{j,h}$ is the time interval from the time that $P_{j,h-1}$ is transmitted to the receiver to the time that $P_{j,h}$ is transmitted to the receiver. Here, we note that if $h = 1$, $P_{j,0} = P_{j-1}$. We find that H_j follows a geometric distribution, where the PMF of H_j is given by

$$\Pr(H_j = H) = (1-\varepsilon_\gamma)\,\varepsilon_\gamma^{H-1}. \tag{4.44}$$

In (4.43), $B_{j,h}$ is calculated by

$$B_{j,h} = V_{j,h} + M, \tag{4.45}$$

where $V_{j,h}$ is the time that the transmitter is in the idle state during $B_{j,h}$. Since $V_{j,k}$ follows an exponential distribution, we obtain

$$\mathbb{E}\left[e^{\alpha B_{j,h}}\right] = e^{\alpha M}\mathbb{E}\left[e^{\alpha V_{j,h}}\right] = e^{\alpha M}\frac{\lambda}{\lambda-\alpha}. \tag{4.46}$$

By substituting (4.46) and (4.44) into (4.43), we obtain $\mathbb{E}\left[e^{\alpha\zeta_j}\right]$ as

$$\mathbb{E}\left[e^{\alpha\zeta_j}\right] = \sum_{h=1}^{\infty}(1-\varepsilon_\gamma)\,\varepsilon_\gamma^{h-1}\mathbb{E}\left[e^{\alpha B_{j,h}}\right]^h = \frac{\lambda\,(1-\varepsilon_\gamma)\,e^{\alpha M}}{\lambda(1-\varepsilon_\gamma e^{\alpha M})-\alpha}. \tag{4.47}$$

Finally, based on (4.43) and (4.45), we calculate $\mathbb{E}[\zeta_j]$ as

$$\mathbb{E}[\zeta_j] = \frac{\lambda M + 1}{\lambda\,(1-\varepsilon_\gamma)}. \tag{4.48}$$

By substituting (4.42), (4.47), and (4.48) into (4.40), we obtain the final result in (4.41), which completes the proof. ■

We find from Theorem 4.4 that in order to guarantee the existence of a finite average AoI penalty, α needs to satisfy two conditions, namely, $\alpha < -\frac{\log\varepsilon_\gamma}{M}$ and $\lambda\left(1-\varepsilon_\gamma e^{\alpha M}\right) > \alpha$. Otherwise, the average AoI penalty goes to infinity in the NPNB strategy.

We then derive and present the closed-form expression for the average AoI penalty achieved by the NPOB strategy in the following theorem.

Theorem 4.5. Considering the NPOB strategy with the received SNR, γ, the closed-form expression for the average AoI penalty is derived as

$$\overline{C}_{\text{NPOB}} = \frac{\beta\lambda e^{\alpha M}\left(1-\varepsilon_\gamma\right)}{\alpha\left(\lambda M + e^{-\lambda M}\right)}\left(e^{-\lambda M}+\frac{\lambda}{\lambda-\alpha}\left(1-e^{(\alpha-\lambda)M}\right)\right)$$
$$\times\left(\frac{\left(1-\varepsilon_\gamma\right)e^{\alpha M}\left(\lambda-\alpha(1-e^{-\lambda M})\right)}{\lambda\left(1-e^{\alpha M}\varepsilon_\gamma\right)-\alpha\left(1-\varepsilon_\gamma(e^{\alpha M}-e^{(\alpha-\lambda)M})\right)}-1\right)-\beta. \tag{4.49}$$

Proof: Again, we need to derive $\mathbb{E}\left[e^{\alpha T_{j-1}}\right]$, $\mathbb{E}\left[e^{\alpha\zeta_j}\right]$, and $\mathbb{E}[\zeta_j]$ to obtain the average AoI penalty. We first derive $\mathbb{E}\left[e^{\alpha T_{j-1}}\right]$ as

$$\mathbb{E}\left[e^{\alpha T_{j-1}}\right] = \mathbb{E}\left[e^{\alpha(M+W_{j-1})}\right] = e^{\alpha M}\mathbb{E}\left[e^{\alpha W_{j-1}}\right], \tag{4.50}$$

where W_{j-1} is the waiting time of P_{j-1} in the buffer. We note that $W_{j-1}=0$ when P_{j-1} is generated after the transmission of P_{j-2} and it follows an exponential distribution when P_{j-1} is generated during the transmission of P_{j-2}. We then calculate $\mathbb{E}\left[e^{\alpha W_{j-1}}\right]$ as

$$\mathbb{E}\left[e^{\alpha W_{j-1}}\right] = \int_0^M e^{\alpha t}\lambda e^{-\lambda t}\mathrm{d}t + \int_M^\infty \lambda e^{-\lambda t}\mathrm{d}t = e^{-\lambda M}+\frac{\lambda}{\lambda-\alpha}\left(1-e^{(\alpha-\lambda)M}\right). \tag{4.51}$$

By substituting (4.51) into (4.50), we obtain $\mathbb{E}[e^{\alpha T_{j-1}}]$.

We then derive $\mathbb{E}\left[e^{\alpha\zeta_j}\right]$ and $\mathbb{E}[\zeta_j]$. We note that $V_{j,h}$ exists if and only if $P_{j,h}$ is generated when the transmitter is in the idle state after $P_{j,h-1}$ is served and the probability of $V_{j,h}=0$ is $\Pr\left(V_{j,h}=0\right)=1-e^{-\lambda M}$ for the NPOB strategy. Hence, we calculate $\mathbb{E}\left[e^{\alpha V_{j,h}}\right]$ and $\mathbb{E}[V_{j,h}]$ as

$$\mathbb{E}\left[e^{\alpha V_{j,h}}\right] = 1-e^{-\lambda M}+e^{-\lambda M}\int_0^\infty e^{\alpha t}\lambda e^{-\lambda t}\mathrm{d}t = 1+\frac{\alpha e^{-\lambda M}}{\lambda-\alpha}. \tag{4.52}$$

and

$$\mathbb{E}\left[V_{j,h}\right] = e^{-\lambda M}\int_0^\infty t\lambda e^{-\lambda t}\mathrm{d}t = \frac{e^{-\lambda M}}{\lambda}, \tag{4.53}$$

respectively. By substituting (4.52) and (4.53) into (4.45), and combining the result with (4.43), we obtain $\mathbb{E}\left[e^{\alpha\zeta_j}\right]$ as

$$\mathbb{E}\left[e^{\alpha\zeta_j}\right] = \frac{(1-\varepsilon_\gamma)e^{\alpha M}\left(\lambda-\alpha(1-e^{-\lambda M})\right)}{\lambda\left(1-e^{\alpha M}\varepsilon_\gamma\right)-\alpha\left(1-\varepsilon_\gamma(e^{\alpha M}-e^{(\alpha-\lambda)M})\right)} \tag{4.54}$$

and obtain $\mathbb{E}\left[\zeta_j\right]$ as

$$\mathbb{E}\left[\zeta_j\right] = \frac{\lambda M + e^{-\lambda M}}{\lambda\left(1-\varepsilon_\gamma\right)}. \tag{4.55}$$

By substituting (4.50), (4.54), and (4.55) into (4.40), we obtain the final result in (4.49), completing this proof. ■

We find from Theorem 4.5 that in order to guarantee the existence of a finite average AoI penalty, α needs to satisfy two conditions, namely, $\alpha < -\frac{\log \varepsilon_\gamma}{M}$ and $(\lambda-\alpha)\left(1-\varepsilon_\gamma e^{\alpha M}\right) > \alpha \varepsilon_\gamma e^{(\alpha-\lambda)M}$. Otherwise, the average AoI penalty goes to infinity in the NPOB strategy.

We note that the information freshness is usually improved by eliminating the waiting time of the transmitter for a new packet and the waiting time of a packet in the buffer, which is the zero-waiting policy. Under this policy, i.e., $\lambda \to \infty$, we find that the average AoI penalty achieved by the NPNB strategy is same as the average AoI penalty achieved by the NPOB strategy, which is given by

$$\overline{C}_{\mathrm{ZW}} = \frac{\beta e^{\alpha M}\left(1-\varepsilon_\gamma\right)\left(e^{\alpha M}-1\right)}{\alpha M\left(1-\varepsilon_\gamma e^{\alpha M}\right)} - \beta, \tag{4.56}$$

where $\alpha < -\frac{\log \varepsilon_\gamma}{M}$ needs to be satisfied to guarantee the existence of a finite average AoI penalty.

We further derive the average AoI penalty achieved by the preemption strategy in the following Theorem.

Theorem 4.6. Considering the preemption strategy with the received SNR, γ, the closed-form expression for the average AoI penalty is derived as

$$\overline{C}_{\mathrm{P}} = \frac{\beta\lambda e^{(\alpha-\lambda)M}\left(1-\varepsilon_\gamma\right)\left(\alpha e^{\lambda M}-\lambda\varepsilon_\gamma\left(e^{\alpha M}-1\right)\right)}{\alpha\left(\left(\lambda e^{\alpha M}-\alpha e^{\lambda M}\right)-\lambda\varepsilon_\gamma\right)} - \beta. \tag{4.57}$$

Proof: To obtain the average AoI penalty, $\mathbb{E}\left[e^{\alpha T_{j-1}}\right]$, $\mathbb{E}\left[e^{\alpha\zeta_j}\right]$, and $\mathbb{E}\left[\zeta_j\right]$ need to be derived. Since $T_{j-1} = M$ for the preemption strategy, we obtain $\mathbb{E}\left[e^{\alpha T_{j-1}}\right]$ as

$$\mathbb{E}\left[e^{\alpha T_{j-1}}\right] = e^{\alpha M}. \tag{4.58}$$

We then derive $\mathbb{E}\left[e^{\alpha\zeta_j}\right]$ and $\mathbb{E}\left[\zeta_j\right]$. For the preemption strategy, we note that

$$B_{j,h} = V_{j,h} + \sum_{k=0}^{K_{j,h}} S_{j,h,k} + M, \tag{4.59}$$

where $K_{j,h}$ is the preemption times of the packet of $P_{j,h}$ and $S_{j,h,k}$ is the transmission time of the kth preempted packet. We note that $P_{j,h}$ is not preempted, i.e., $K_{j,h} = 0$, with the probability $\Pr(K_{j,h} = 0) = e^{-\lambda M}$ and $K_{j,h}$ follows a geometric distribution, i.e., $\Pr(K_{j,h} = K) = (1-$

$e^{-\lambda M})^{K} e^{-\lambda M}$. Hence, we calculate $\mathbb{E}\left[e^{\alpha B_{j,h}}\right]$ and $\mathbb{E}\left[B_{j,h}\right]$ as

$$\mathbb{E}\left[e^{\alpha B_{j,h}}\right] = \frac{\lambda e^{\alpha M}}{\lambda - \alpha} \sum_{k=0}^{\infty} e^{-\lambda M} \left(\frac{\lambda \left(e^{\lambda M} - e^{\alpha M}\right)}{\left(\lambda - \alpha\right) e^{\lambda M}} \right)^{k} = \frac{\lambda e^{\alpha M}}{\lambda e^{\alpha M} - \alpha e^{\lambda M}}, \tag{4.60}$$

and

$$\mathbb{E}\left[B_{j,h}\right] = \frac{1}{\lambda} + M + \mathbb{E}\left[K_{j,h}\right] \mathbb{E}\left[S_{j,h,k}\right] = \frac{1}{\lambda e^{-\lambda M}}, \tag{4.61}$$

respectively. By substituting (4.60) and (4.61) into (4.43), we obtain $\mathbb{E}\left[e^{\alpha \zeta_j}\right]$ and $\mathbb{E}\left[\zeta_j\right]$ as

$$\mathbb{E}\left[e^{\alpha \zeta_j}\right] = \frac{\lambda e^{\alpha M}(1 - \varepsilon_\gamma)}{\lambda e^{\alpha M} - \alpha e^{\lambda M} - \lambda \varepsilon_\gamma}, \tag{4.62}$$

and

$$\mathbb{E}\left[\zeta_j\right] = \frac{1}{\lambda e^{-\lambda M} \left(1 - \varepsilon_\gamma\right)}, \tag{4.63}$$

respectively. By substituting (4.58), (4.62), and (4.63) into (4.40), we obtain the final result in (4.57). This completes the proof. ■

We find from Theorem 4.6 that in order to guarantee the existence of a finite average AoI penalty, α needs to satisfy two conditions, namely, $\alpha < \lambda$ and $\left(\lambda e^{\alpha M} - \alpha e^{\lambda M}\right) > \lambda \varepsilon_\gamma$. Otherwise, the average AoI penalty goes to infinity in the preemption strategy.

So far, we have obtained the expressions for the average AoI penalty achieved by the NPNB strategy, the NPOB strategy, and the preemption strategy in Theorems 4.4, 4.5, and 4.6, respectively. Since the block error rate is the function of the packet blocklength, there would exist the optimal coding rate to minimize the average AoI penalty for three strategies. Due to the complex form of the block error rate given in (4.34), it is difficult to derive closed-form expressions for the optimal packet blocklength for these three strategies. Thus, we resort to numerical methods, e.g., numerical search methods, to find the optimal packet blocklengths.

4.3.3 Numerical Results

In this section, we present numerical results to evaluate the impact of various parameters, including the coding rate, status update generation rate, and the value of α on the average AoI penalty, achieved by three strategies in our considered system. We then examine how the value of α affects the optimal coding rate in the system.

Fig. 4.6 plots the average AoI penalty versus the coding rate, R. We first observe that the analytical average AoI penalty precisely matches the simulation results, which demonstrates

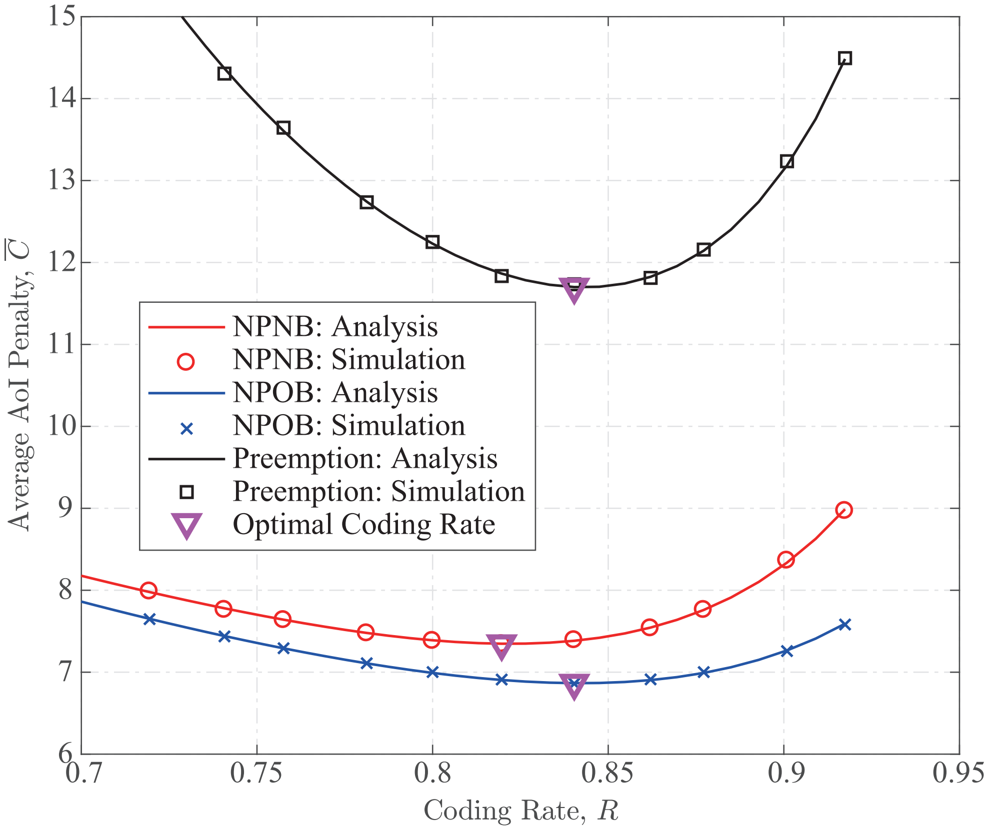


Figure 4.6: Average AoI penalty versus the coding rate, R, with $\lambda = 0.01$, $\gamma = 3$, and $L = 100$ for $C(\Delta(t)) = 10\left(e^{0.002\Delta(t)} - 1\right)$.

the correctness of our analytical results in Theorems 4.4, 4.5, and 4.6. We then observe that for three strategies, the average AoI penalty first decreases and then increases when R increases. This observation is due to the fact that the increase in R has a two-fold effect on the average AoI penalty through the packet blocklength and the block error rate. When R is small, the increase in R leads to the smaller packet blocklength, which decreases the average AoI penalty. When R is large and exceeds a certain threshold, the increase in R brings about a significant increase in the block error rate, thereby increasing the average AoI penalty.

Fig. 4.7 plots the average AoI penalty versus the status update generation rate, λ. We first observe that for both the NPNB and NPOB strategies, the average AoI penalty decreases monotonically when λ increases, which approaches the average AoI penalty under the zero-waiting policy when λ is large. This is because the increase in λ decreases the transmitter's waiting time for a new status update, which in turn decreases the average AoI penalty. We then observe that the NPOB strategy achieves a lower average AoI penalty than the NPNB strategy when λ is small, but a larger average AoI penalty when λ is large. This is because when λ is small, the NPOB strategy has a higher update frequency, which deceases the average AoI

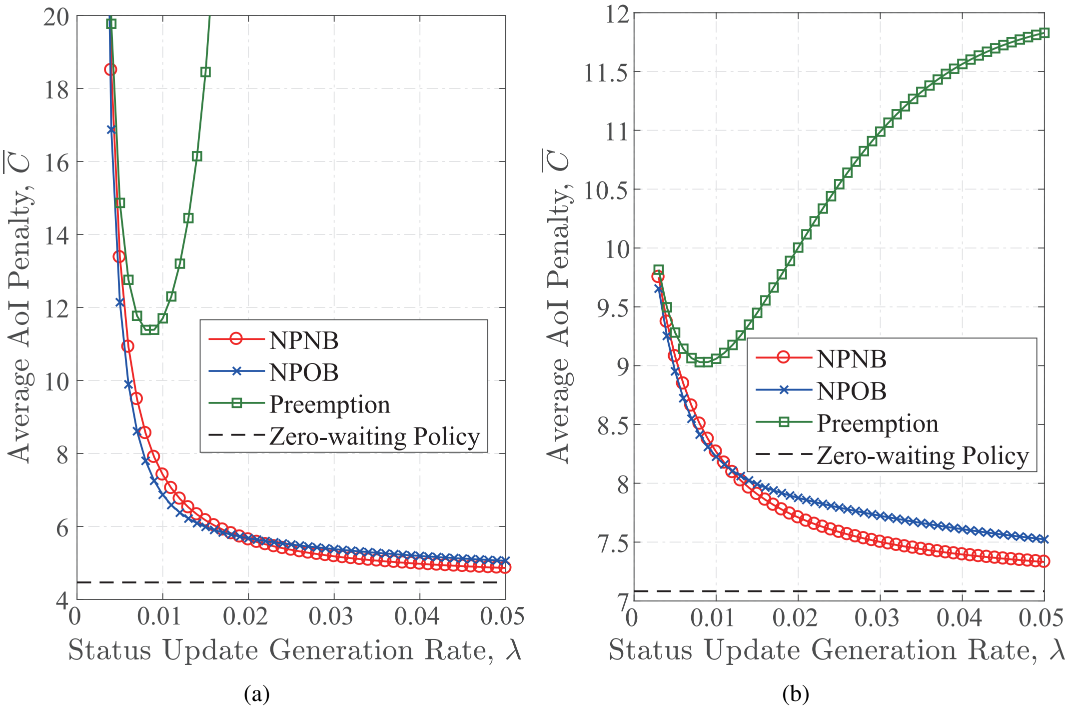


Figure 4.7: Average AoI penalty versus the status update generation rate, λ, with $R = 0.85$, $\gamma = 3$, and $L = 100$, for (a) $C(\Delta(t)) = 10\left(e^{0.002\Delta(t)} - 1\right)$ and (b) $C(\Delta(t)) = -12\left(e^{-0.005\Delta(t)} - 1\right)$.

penalty. When λ is large and exceeds a threshold, the priority of this higher update frequency reduces and the transmitted packets are outdated compared to the NPNB strategy, leading to the higher average AoI penalty. We further observe that the average AoI first decreases and then increases when λ increases in the preemption strategy. This observation is due to the fact that the increase in λ has a two-fold effect on the average AoI penalty. When λ is small, the increase in λ reduces the transmitter's waiting time after a successful transmission, which decreases the average AoI penalty. When λ exceeds a certain threshold, the increase in λ leads to a significant increase in the probability that a status update is preempted, thereby increasing the average AoI penalty. In addition, we find that non-preemption strategies, including the NPNB and NPOB strategies, achieve a better AoI performance than the preemption strategy.

Fig. 4.8 plots the optimal coding rate versus the value of α. We observe that the optimal coding rate decreases as the value of α increases. Here, we note that a lower coding rate represents fewer transmission errors and higher transmission reliability. It implies that the system requires high reliability when having a relatively low tolerance to outdated data, but low reliability when having a relatively high tolerance to outdated data. It follows that the non-linear AoI penalty reflects not only the receiver's dissatisfaction on outdated data but also the reliability requirement of short packet communications. Based on this observation, the value

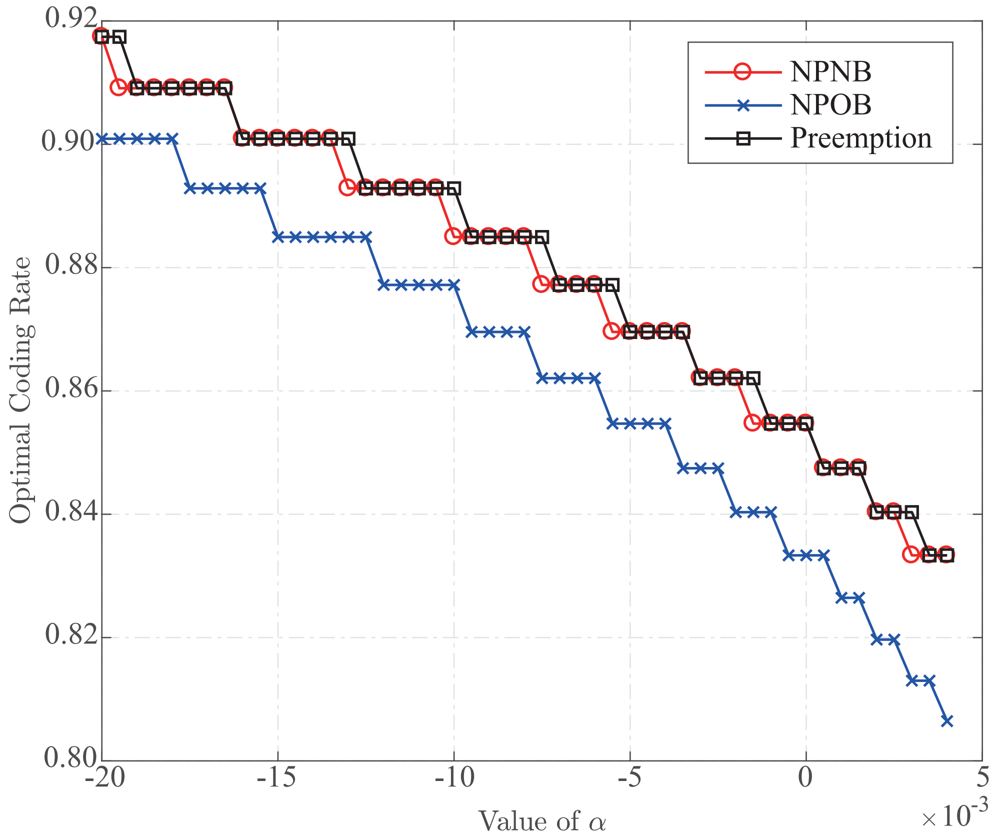


Figure 4.8: The optimal coding rate of the system versus the value of α, with $\lambda = 0.01$, $\gamma = 3$, $L = 100$, and $\beta = \frac{1}{\alpha}$.

of α can be designed according to the transmission reliability requirement.

4.3.4 Summary

This section considered a point-to-point system where a transmitter generates packets and transmits them to a receiver via an unreliable channel. By considering non-linear α-β AoI penalty, we analyzed the average AoI penalty achieved by three packet management strategies, i.e., the NPNB strategy, the NPOB strategy, and the preemption strategy, based on a finite packet blocklength model. With simulation results, we demonstrated the accuracy of our analysis and showed that the NPNB strategy and NPOB strategies are more suitable than the preemption strategy for short packet communication systems by achieving a lower average AoI penalty. Moreover, we found that the system requires high transmission reliability when the receiver has a low tolerance to outdated data, but low transmission reliability when the receiver has a high tolerance.

Chapter 5

AoI Based Transmission Scheme Selection

5.1 Introduction

In this chapter, we analytically address the question of whether to adopt broadcast transmission or unicast transmission for the optimal AoI performance. In particular, we study the AoI performance of a multi-user downlink system where a BS generates and transmits status updates to multiple UEs with two transmission schemes, i.e., the broadcast transmission scheme and the unicast transmission scheme. In the broadcast transmission scheme, the status update for all UEs is jointly encoded into a packet for transmission, while in the unicast transmission scheme, the status update for each UE is encoded individually and transmitted by following the RR policy. For both transmission schemes, we examine three packet management strategies, namely the non-preemption strategy, the preemption in buffer strategy, and the preemption in serving strategy. We first derive new closed-form expressions for the average AoI achieved by two transmission schemes with three packet management strategies. Then, we compare the AoI performance of two transmission schemes in two systems, namely, a remote control system where the distances between the control BS and all UEs are almost same, and a dynamic system where the location of UEs follows a two-dimensional Poison point process (2D-PPP) within the coverage area of the BS. Aided by simulation results, we demonstrate the accuracy of our analytical results and discuss the appropriate choice of the transmission scheme to decrease the average AoI in the considered system. We find that the unicast transmission scheme is more appropriate for the system with a large number of UEs, while the broadcast transmission scheme is more appropriate for the system with a small number of UEs.

This chapter is organized as follows. In Section 5.2, the system model and two transmission schemes are described. New closed-form expressions for the average AoI of two transmission schemes with different packet management strategies are derived in Section 5.3. In Section 5.4, the appropriate transmission schemes are decided based on the AoI performance in two systems. The numerical results are discussed in Section 5.5 and the summary is presented in

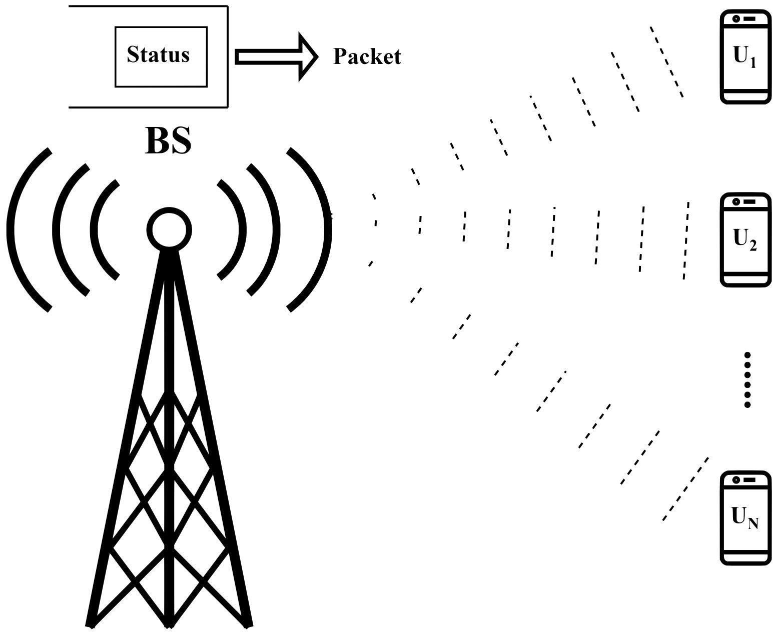


Figure 5.1: Illustration of our considered multi-user system where the BS transmits status update to N UEs.

Section 5.6.

5.2 System Model

5.2.1 System Description

We consider a multi-user wireless communication system, as depicted in Fig. 5.1, where a BS transmits status updates to N UEs[1]. We denote the nth UE by U_n, where $n = 1, 2, \cdots, N$. In this system, the BS generates the status update for N UEs according to a Poisson process with the rate λ. Then, packets are generated by the BS based on the generated status update, and the packets are transmitted from the BS to UEs. In this system, we consider two transmission schemes, namely, the broadcast transmission scheme and the unicast transmission scheme, described as follows:

- In the broadcast transmission scheme, we assume that the status update for N UEs contains L bits, where L is a fixed value. Once the BS generates the status update, it encodes these L bits into a packet with the blocklength of M c.u. and transmits this packet to N UEs. For this transmission, we define the coding rate, R, as the ratio between the number of bits in the status update and the blocklength of the transmitted packet, i.e., $R = \frac{L}{M}$.

[1]Note that the non-feedback mechanism is considered in this chapter to eliminate the feedback overhead.

- In the unicast transmission scheme, the BS separately transmits the status update to N UEs by following an RR policy. We denote the length of the status update and the packet blocklength for U_n by L_n and M_n, respectively. For this transmission, we define the coding rate for U_n, R_n, as the ratio between the number of bits in the status update and the blocklength of the transmitted packet, i.e., $R_n = \frac{L_n}{M_n}$.

We clarify that the total number of bits of the status update for N UEs is L in the broadcast transmission scheme and $\sum_{n=1}^{N} L_n$ in the unicast transmission scheme. According to the information correlation among UEs, the joint information is less than or equal to the sum of individual information, i.e., $L \leq \sum_{n=1}^{N} L_n$. Thus, we denote α as the information ratio between the broadcast transmission scheme and the unicast transmission scheme and express it by $\alpha = \frac{L}{\sum_{n=1}^{N} L_n}$, where $0 < \alpha \leq 1$.

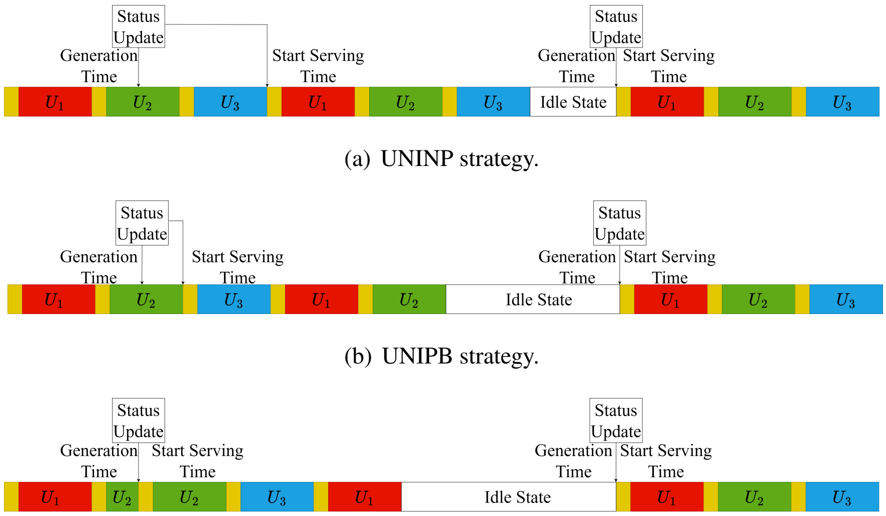
We assume an AWGN channel between the BS and each UE, which has been widely considered in the literature, e.g., [63, 64, 67, 102]. According to [62], the block error rate for U_n using finite block length coding can be approximated as

$$\varepsilon_n(l,m,\gamma_n) = Q\left(\frac{\frac{1}{2}\log_2(1+\gamma_n) - \frac{l}{m}}{\log_2(e)\sqrt{\frac{1}{2m}\left(1 - \frac{1}{(1+\gamma_n^2)}\right)}}\right), \tag{5.1}$$

where l is the number of bits in the status update, m is the packet blocklength, and γ_n is the received SNR at U_n. We clarify that the results obtained for the AWGN channel can be extended to the small-scale fading channel[2]. We note that (5.1) is very tight for short packet communications when $m \geq 100$ [62]. For our considered system, we denote T_u as the transmission time for each c.u.. To facilitate our analysis, we define T_u as the unit time.

We note that, apart from the transmission time, the BS requires an extra pre-processing and status update processing time to send connection requests for establishing the transmission link with each UE and processing the status update for each UE in the unicast transmission scheme, compared with the broadcast transmission scheme [116]. Here, we denote M_{L} as the time spent on the pre-processing and status update processing for each UE. Hence, we define the serving time as $M_{\mathrm{L}} + M_n$ and the ratio between the transmission time and the serving time for U_n is $\rho_n = \frac{M_n}{M_{\mathrm{L}}+M_n}$ in the unicast transmission scheme.

[2]With small-scale fading, the block error rate for U_n using finite block length coding can be approximated as $\varepsilon_n(l,m,\gamma_n) = \int_{\hbar_n=0}^{\infty} f(\hbar_n) Q\left(\frac{\frac{1}{2}\log_2(1+\gamma_n\hbar_n) - \frac{l}{m}}{\log_2(e)\sqrt{\frac{1}{2m}\left(1-\frac{1}{(1+(\gamma_n\hbar_n)^2)}\right)}}\right) \mathrm{d}\hbar_n$, where $\hbar_n$ is the small-scale channel gain of U_n and $f(\hbar_n)$ is the PDF of $\hbar_n$.



(a) UNINP strategy.

(b) UNIPB strategy.

(c) UNIPS strategy.

Figure 5.2: An example to illustrate the timeliness of the unicast transmission scheme with three packet management strategies in a three-UE system.

5.2.2 Packet Management Strategies

To ensure the freshness of packets, we assume that the LCFS queuing policy is adopted and the buffer at the BS only stores the newest status update. In the broadcast transmission scheme, when a new status update is generated and the BS is in the idle state, the BS encodes the status update into a packet and transmits this packet to all UEs. When a new status update is generated at the BS and the BS is transmitting a packet, we examine two packet management strategies, namely, the non-preemption strategy and the preemption in serving strategy, detailed as follows:

- Broadcast transmission with non-preemption (BRNP) strategy: In this strategy, if a new status update is generated at the BS and the BS is transmitting a packet, the BS stores the new status update in the buffer and keeps the current transmission.
- Broadcast transmission with preemption in serving (BRPS) strategy: In this strategy, preemption is considered such that the newly generated status update always preempts the transmission of the current status update. Specifically, if a new status update is generated at the BS and the BS is transmitting a status update, the BS discards the current transmission and starts to transmit the new status update.

In the unicast transmission scheme, when a new status update is generated and the BS is in the idle state, the BS processes the status update and transmits the packets to N UEs, i.e., U_1, U_2, $\cdots$, U_N. In particular, the BS processes the status update and encodes the required status update information for U_n into a packet. Then the BS establishes the transmission link with U_n and transmits this packet. After the transmission of U_n's packet, the BS repeats this step for U_{n+1} until all UEs are served. We note that the BS sends the status update only once for each UE. When a new status update is generated at the BS and the BS is serving a packet, we examine three packet management strategies, namely the non-preemption strategy, the preemption in buffer strategy, and the preemption in serving strategy, detailed as follows:

- Unicast transmission with non-preemption (UNINP) strategy: In this strategy, if a new status update is generated at the BS and the BS is serving a UE, the BS stores the new status update in the buffer and keeps the current service. We note that the serving order of each status update is always from U_1 to U_N in this strategy.

- Unicast transmission with preemption in buffer (UNIPB) strategy: In this strategy, when a new status update is generated at the BS and the BS is serving U_n, the BS stores the new status update in the buffer. After the transmission of U_n's packet, the BS processes this status update and transmits it to U_{n+1}.

- Unicast transmission with preemption in serving (UNIPS) strategy: In this strategy, preemption is considered such that the newly generated status update always preempts the service of the current status update. Specifically, if a new status update is generated at the BS and the BS is serving U_n, the BS discards the current serving and starts to serve the new status update for U_n.

In Fig 5.2, we provide an example of the timeliness of the unicast transmission scheme with three packet management strategies in a three-UE system. We note that the BRNP strategy and the BRPS strategy can be considered as the special case of the UNINP strategy and the UNIPS strategy with $N = 1$ and $M_{\mathrm{L}} = 0$, respectively.

5.2.3 Formulation of AoI

In this subsection, we formulate the average AoI of the considered system. Without loss of generality, we arbitrarily select one UE, U_n, and analyze its average AoI, Δ_n. We denote $\Delta_n(t)$ as the AoI of U_n at time t. We assume that the observation begins at $t = 0$ with the AoI of $\Delta_n(0)$. Then, we express the AoI at time t as

$$\Delta_n(t) = t - u_n(t), \tag{5.2}$$

where $u_n(t)$ is the generation time of the most recently received status update at U_n at time t. Then, the time-average AoI of U_n over the observation time interval $(0, \tau)$ is calculated as

$$\Delta_n = \frac{1}{\tau}\int_0^{\tau} \Delta_n(t)\mathrm{d}t. \tag{5.3}$$

We denote P_j as the jth successfully received status update for U_n generated after time $t=0$, $j=1,2,\cdots$. We then denote ζ_j as the time interval from the received time of P_{j-1} to the received time of P_j and denote T_j as the time interval from the generation time of P_j to the received time of P_j. Therefore, we express ζ_j and T_j as

$$\zeta_j = t'_j - t'_{j-1} \tag{5.4}$$

and

$$T_j = t'_j - t_j, \tag{5.5}$$

respectively, where t_j is the generation time of P_j and t'_j is the time for U_n to receive P_j. According to [19], the average AoI of U_n is thus calculated as

$$\Delta_n = \frac{\mathbb{E}[\zeta_j^2]}{2\mathbb{E}[\zeta_j]} + \mathbb{E}[T_j], \tag{5.6}$$

where $\mathbb{E}[\cdot]$ denotes the expectation. By averaging Δ_n over all UEs, we obtain the average AoI of the considered system as

$$\Delta = \frac{1}{N}\sum_{n=1}^{N}\Delta_n. \tag{5.7}$$

5.3 Closed-form Analysis of Average AoI

In this section, we derive new closed-form expressions for the average AoI of U_n in two transmission schemes with different packet management strategies. We first focus on the UNINP strategy and derive the closed-form expression for its average AoI in the following Theorem.

Theorem 5.1. In the UNINP strategy, the closed-form expression for the average AoI of U_n with the received SNR, γ_n, is derived as

$$\begin{aligned}\Delta_{\text{UNINP},n} =& \frac{1+\varepsilon_n}{2(1-\varepsilon_n)}\left(M_{\text{T}} + \frac{1}{\lambda}e^{-\lambda M_{\text{T}}}\right) + \frac{(2e^{-\lambda M_{\text{T}}} - e^{-2\lambda M_{\text{T}}})}{2(\lambda^2 M_{\text{T}} + \lambda e^{-\lambda M_{\text{T}}})} \\ &- \sum_{k=n+1}^{N} M'_k + \left(\frac{1}{\lambda} + M_{\text{T}}\right)(1-e^{-\lambda M_{\text{T}}}),\end{aligned} \tag{5.8}$$

where $M'_k = M_k + M_{\text{L}}$, $M_{\text{T}} = \sum_{k=1}^{N} M'_k$, and $\varepsilon_n = \varepsilon(L_n, M_n, \gamma_n)$.

Proof: See Appendix B.1. ■

Using Theorem 5.1, the average AoI of the system in the UNINP strategy can be obtained by averaging Δ_n over all UEs. We find that the average AoI of the system is minimized if M_n

is monotonically non-decreasing w.r.t. n, i.e., $M_{n_1} \geq M_{n_2}$, for any $n_1 > n_2$. It indicates that the UE with the shortest blocklength needs to be served first in the UNINP strategy.

We next derive the average AoI in the UNIPB strategy. To facilitate our derivation, we define $g_{k_1,k_2} = k_1 + k_2$ if $k_1 + k_2 \leq N$, but $g_{k_1,k_2} = k_1 + k_2 - N$ if $k_1 + k_2 > N$, where k_1 and $k_2 = 1, 2, \cdots, N$. We then present the closed-form expression for the average AoI in the following theorem.

Theorem 5.2. In the UNIPB strategy, the closed-form expression for the average AoI of U_n with the received SNR, γ_n, is derived as

$$\begin{aligned}\Delta_{\text{UNIPB},n} =& \frac{1+\varepsilon_n}{2(1-\varepsilon_n)}(M_\text{T}+\xi) + \frac{1-e^{-\lambda M_\text{T}}}{\lambda} + \frac{2\xi - \lambda\xi^2}{2\lambda(M_\text{T}+\xi)} \\ & - \sum_{k=1}^{N} M'_{g_{k,n}} \left(\frac{e^{-\lambda M_\text{T}}\left(1-e^{-\lambda \sum_{\kappa=k}^{N-1} M'_{g_{\kappa,n}}}\right)}{1-e^{-\lambda M_\text{T}}} \right) + M'_n(1-e^{-\lambda M_\text{T}}),\end{aligned} \tag{5.9}$$

where $\xi = \frac{e^{-\lambda M_\text{T}}}{\lambda(1-e^{-\lambda M_\text{T}})} \sum_{k=1}^{N} (1-e^{-\lambda M'_{g_{k,n}}})$.

Proof: See Appendix B.2. ■

To improve the freshness of information, the system is always designed to eliminate the waiting time in the buffer, which is the zero-waiting policy. Under this policy, we obtain the closed-form expressions for the average AoI in the UNINP and UNIPB strategies in the following Corollary.

Corollary 5.1. With the received SNR, γ_n, the closed-form expression for the average AoI of U_n in the UNINP strategy under the zero-waiting policy is given by

$$\Delta_{\text{UNINPZ},n} = \frac{1+\varepsilon_n}{2(1-\varepsilon_n)} M_\text{T} + M_\text{T} - \sum_{k=n+1}^{N} M'_k, \tag{5.10}$$

and the closed-form expression for the average AoI of U_n in the UNIPB strategy under the zero-waiting policy is given by

$$\Delta_{\text{UNIPBZ},n} = \frac{1+\varepsilon_n}{2(1-\varepsilon_n)} M_\text{T} + M'_n. \tag{5.11}$$

Proof: Since the LCFS queuing policy is adopted in the BRNP and UNINP strategies, the zero-waiting policy is achieved when $\lambda \to \infty$. Hence, we obtain $\Delta_{\text{UNINPZ},n} = \lim\limits_{\lambda\to\infty} \Delta_{\text{UNINP},n}$ and $\Delta_{\text{UNIPBZ},n} = \lim\limits_{\lambda\to\infty} \Delta_{\text{UNIPB},n}$, resulting in (5.10) and (5.11), respectively. ■

From Corollary 5.1, we find that the UNIPB strategy always outperforms the UNINP strategy by achieving a lower average AoI under the zero-waiting policy. This is due to the fact that comparing with the UNINP strategy, the UNIPB strategy eliminates the waiting time of status update for each UE caused by the service of other UEs.

We further derive the average AoI in the UNIPS strategy in the following Theorem.

Theorem 5.3. In the UNIPS strategy, the closed-form expression for the average AoI of U_n with the received SNR, γ_n, is derived as

$$\begin{aligned}\Delta_{\text{UNIPS},n} = & \frac{1+\varepsilon_n}{2\lambda(1-e^{-\lambda M_\text{T}})(1-\varepsilon_n)} \sum_{k=1}^{N} \frac{p_k}{1-p_k} + \sum_{k=1}^{N} M'_{g_{k,n}} \frac{e^{-\lambda \sum_{\kappa=k+1}^{N} M'_{g_{\kappa,n}}} - e^{-\lambda M_\text{T}}}{1-e^{-\lambda M_\text{T}}} \\ & + \frac{\lambda e^{-\lambda M_\text{T}}}{2\psi} \left[\frac{2\psi-\psi^2}{\lambda^2} + \sum_{k=1}^{N} \left(\frac{p_k}{1-p_k} \left(\frac{2}{\lambda^2} + \frac{p_k}{\lambda^2(1-p_k)} \right) - \frac{2M'_{g_{k,n}}}{\lambda(1-p_k)} \right) \right. \\ & \left. - \frac{2e^{-\lambda M_\text{T}}}{1-e^{-\lambda M_\text{T}}} \sum_{k=1}^{N} \frac{p_k}{1-p_k} \sum_{\kappa=1}^{k-1} \frac{p_\kappa}{1-p_\kappa} \left(\frac{1}{\lambda} - \frac{M'_{g_{\kappa,n}}(1-p_\kappa)}{p_\kappa} \right) \right], \end{aligned} \tag{5.12}$$

where $p_k = 1 - e^{-\lambda M'_{g_{k,n}}}$ and $\psi = \frac{e^{-\lambda M_\text{T}}}{1-e^{-\lambda M_\text{T}}} \sum_{k=1}^{N} \frac{p_k}{1-p_k}$.

Proof: See Appendix B.3. ■

We clarify that in Theorem 5.1–5.3 and Corollary 5.1, the transmission time, $M_n = L_n/R_n$, is fixed when the length of the status update, L_n, and the coding rate, R_n, for U_n are fixed.

Since the BRNP strategy and the BRPS strategy are the special case of the UNINP strategy and the UNIPS strategy with $N = 1$ and $M_\text{L} = 0$, respectively, we obtain the closed-form expressions for the average AoI in the BRNP strategy and the BRPS strategy in the following Corollary.

Corollary 5.2. The closed-form expressions for the average AoI of U_n with the received SNR, γ_n, in the BRNP strategy and the BRPS strategy are given by

$$\Delta_{\text{BRNP},n} = \frac{1+\varepsilon_n}{2(1-\varepsilon_n)} \left(M + \frac{1}{\lambda} e^{-\lambda M} \right) + \frac{2e^{-\lambda M} - e^{-2\lambda M}}{2\left(\lambda^2 M + \lambda e^{-\lambda M}\right)} + \left(\frac{1}{\lambda} + M \right) \left(1 - e^{-\lambda M} \right) \tag{5.13}$$

and

$$\Delta_{\text{BRPS},n} = \frac{1}{\lambda e^{-\lambda M}(1-\varepsilon_n)}, \tag{5.14}$$

respectively.

Proof: By substituting $L = L_1$, $M = M_1$, $N = 1$, and $M_\text{L} = 0$ into (5.8) and (5.12), we obtain the closed-form expressions for the average AoI of U_n in the BRNP strategy and the BRPS strategy, given by (5.13) and (5.14), respectively. ■

In Corollary 5.2, the transmission time, $M = L/R$, is fixed when the length of the status update, L, and the coding rate, R, are fixed. From Corollary 5.2, we find that the average

AoI monotonically decreases w.r.t. the status update generation rate, λ, in the BRNP strategy. Moreover, we find that the optimal status update generation rate is $\lambda = \frac{1}{M}$ in the BRPS strategy.

5.4 Transmission Scheme Selection

In this section, we investigate the selection of appropriate transmission scheme for two systems, namely, a remote control system and a dynamic system, which correspond to a remote control automation factory and a vehicular network, respectively. To facilitate this investigation, we compare the average AoI achieved by both transmission schemes with the non-preemption strategy, i.e., the BRNP strategy and the UNINP strategy, against the information ratio, α. Specifically, we assume that all UEs require the same length of the status update, i.e., $L_n = L_h$, $n = 1, 2, \cdots, N$.

5.4.1 Remote Control System

In a remote control automation factory, a large number of UEs are employed and the distances between the control BS and these UEs are almost same, since the UEs are located close to each other. Moreover, these UEs may require to cooperate with each other [117], implying that the required information between the cooperated UEs is highly correlated. To characterize these properties in a remote control automation factory, we consider a remote control system in this subsection and compare the average AoI achieved by the BRNP strategy and the UNINP strategy against the information ratio, α. In this system, we assume that UEs have the same received SNR, i.e., $\gamma_n = \gamma_h$, and consider the same coding rate, R, in these two strategies, i.e., $R = \frac{L}{M} = \frac{L_h}{M_h}$, where $M_n = M_h$ and $\rho = \rho_n$, $n = 1, 2, \cdots, N$. We also assume that the coding rate is low, leading to a low block error rate, which is due to the fact that high reliable communication scenarios are of great interests in practice. Theorem 5.4 characterizes the condition of α where the average AoI in one strategy is better than another.

Theorem 5.4. In the remote control system, the BRNP strategy has a better AoI performance than the UNINP strategy, if $\alpha \leq \alpha_{\mathrm{th}}$, and the UNINP strategy has a better AoI performance than the BRNP strategy, if $\alpha > \alpha_{\mathrm{th}}$. Specifically, the information ratio threshold α_{th} satisfies

$$\left(\frac{3}{2} - e^{-\lambda \alpha_{\mathrm{th}} \rho M_{\mathrm{T}}}\right) \alpha_{\mathrm{th}} \rho M_{\mathrm{T}} + \frac{1}{2\lambda} \Omega\left(\lambda \alpha_{\mathrm{th}} \rho M_{\mathrm{T}}\right) = \left(\frac{2N+1}{2N} - e^{-\lambda M_{\mathrm{T}}}\right) M_{\mathrm{T}} + \frac{1}{2\lambda} \Omega\left(\lambda M_{\mathrm{T}}\right), \tag{5.15}$$

where $\Omega(\omega) = \frac{2e^{-\omega} - e^{-2\omega}}{\omega + e^{-\omega}} - e^{-\omega}$.

Proof: See Appendix B.4. ■

Theorem 5.4 reveals that the BRNP strategy is better when there is high correlation among the information for UEs, but the UNINP strategy is better when there is low correlation among the information for UEs. Moreover, we find that the threshold α_{th} monotonically decreases as the ratio between the transmission time and the serving time, ρ, increases. This indicates that the BRNP strategy has a better AoI performance than the UNINP strategy when the time spent on the pre-processing and status update processing, M_{L}, is long.

We then examine the value of α_{th} under the zero-waiting policy, i.e., $\lambda \to \infty$, and the sporadic status update generation rate, i.e., $\lambda \to 0$, in the following Corollary.

Corollary 5.3. The information ratio threshold α_{th} under the zero-waiting policy is given by

$$\alpha_{\text{th}} = \frac{2N+1}{3N\rho}, \tag{5.16}$$

and α_{th} under the sporadic status update generation rate is given by

$$\alpha_{\text{th}} = \frac{N+1}{2N\rho}. \tag{5.17}$$

Proof: When $\lambda \to \infty$, (5.15) can be rewritten as

$$\frac{3}{2}\alpha_{\text{th}}\rho M_{\text{T}} = \frac{2N+1}{2N}M_{\text{T}}, \tag{5.18}$$

which leads to (5.16). When $\lambda \to 0$, we substitute $e^{-\lambda M_{\text{T}}} = 1 - \lambda M_{\text{T}}$ into (5.15), which gives

$$\frac{1}{2}\alpha_{\text{th}}\rho M_{\text{T}} - \frac{1-\lambda\alpha_{\text{th}}\rho M_{\text{T}}}{2\lambda} + \frac{1}{2\lambda^2} = \frac{1}{2N}M_{\text{T}} - \frac{1-\lambda M_{\text{T}}}{2\lambda} + \frac{1}{2\lambda^2}. \tag{5.19}$$

Hence, we obtain (5.17). ■

From Corollary 5.3, we find the relationship between the number of UEs and the information ratio threshold, i.e., the information ratio threshold monotonically decreases when the number of UEs increases. This finding indicates that the UNINP strategy is better than the BRNP strategy when the number of UEs is large. Moreover, we find that the information ratio threshold is high under the zero-waiting policy, but is low under the sporadic status generation rate. This finding indicates that the BRNP strategy is better for large λ and the UNINP strategy is better for small λ.

5.4.2 Dynamic System

In vehicular wireless networks, the number and locations of UEs may change from time to time. Since nearby vehicles in a vehicular network share very similar surrounding traffic environment, it means that all vehicles may share a large amount of common information, e.g., the road condition, traffic condition, and the locations of surrounding buildings, pedestrians, and vehicles. Moreover, each vehicle requires its own individual information, e.g., velocity requirement, navigation service, and control command at intersections. In this subsection, we consider a dynamic system to characterize the features of a vehicular network [118]. In this system, we assume that the BS is located at the center and covers a circular annulus with inner radius D_1 and outer radius D_2, where $D_2 > D_1 \geq 1$. In this area, we assume that the location of UEs follows a 2D-PPP with the intensity λ_{UE}. In this system, we assume that all UEs have the same common information and each UE has its own individual information with the same length. Here, we denote L_{co} as the number of bits of the common information and L_{id} as the number of bits of the individual information for each UE. Hence, with N UEs in the system, the number of bits of the required status update for each UE is $L_h = L_{\text{co}} + L_{\text{id}}$ and the total number of bits of the status update is $L = L_{\text{co}} + NL_{\text{id}}$. We then denote β as the ratio between the individual information and the common information, i.e., $\beta = \frac{L_{\text{id}}}{L_{\text{co}}}$. We note that the information ratio, α, is given as $\alpha = \frac{N\beta+1}{N(\beta+1)}$ of the system with N UEs. Since the information ratio, α, changes as the number of UEs changes, we compare the AoI performance of two strategies based on the ratio between the individual information and the common information, β, in this system. We further assume that all UEs have the same path loss exponent, $\eta \geq 2$, and the received SNR at the reference distance $d_0 = 1$ meter is $\gamma = \gamma_0$. As per the path loss model, the received SNR of UE with the distance d to the BS is $\gamma = \gamma_0 d^{-\eta}$. Accordingly, we denote $C_0 = \frac{1}{2}\log_2(1+\gamma_0)$ as the channel capacity with the received SNR γ_0, $C_{D_1} = \frac{1}{2}\log_2(1+\gamma_{D_1})$ as the channel capacity with the received SNR $\gamma_{D_1} = \gamma_0 D_1^{-\eta}$, and $C_{D_2} = \frac{1}{2}\log_2(1+\gamma_{D_2})$ as the channel capacity with the received SNR $\gamma_{D_2} = \gamma_0 D_2^{-\eta}$. In the BRNP strategy, the BS transmits the status update with the coding rate C_{D_2}, i.e., $R = C_{D_2}$, which covers the entire area. In the UNINP strategy, we assume that the BS detects the location of UEs and transmits their packets based on their location to minimize the average AoI of the system. Since the BS monitors the covered area in real time, the zero-waiting policy is adopted in this system.

Considering the dynamic system, we analyze and compare the expected average AoI achieved by the BRNP strategy and the UNINP strategy in the following Theorem.

Theorem 5.5. In the dynamic system, the expected average AoI in the BRNP strategy is approximated by

$$\mathbb{E}\left[\Delta_{\text{BRNPZ}}\right] \approx \frac{3}{2C_{D_2}}\left(\left(1-e^{-\Lambda}\right)L_{\text{co}} + \Lambda L_{\text{id}}\right), \tag{5.20}$$

and the expected average AoI in the UNINP strategy is approximated by

$$\mathbb{E}\left[\Delta_{\text{UNINPZ}}\right] \approx \left(\frac{1-e^{-\Lambda}}{2} + \Lambda\right)\left(\frac{L_{\text{co}}+L_{\text{id}}}{C_{\Lambda}} + M_{\text{L}}\right), \tag{5.21}$$

where $\Lambda = \lambda_{UE}\pi(D_2^2 - D_1^2)$ and $C_\Lambda = \frac{(D_2^2-D_1^2)\eta}{2^{2+\frac{4C_0}{\eta}}\ln 2(\mathrm{Ei}(x_1)-\mathrm{Ei}(x_2))}$. In C_Λ, $\mathrm{Ei}(x) = -\int_{-x}^{\infty}\frac{e^{-t}}{t}\mathrm{d}t$ is the exponential integral function, $x_1 = -\frac{4}{\eta}C_{D_1}\ln 2$, and $x_2 = -\frac{4}{\eta}C_{D_2}\ln 2$.

Proof: See Appendix B.5. ■

Based on Theorem 5.5, we compare the expected average AoI achieved by two schemes based on the ratio between the individual information and the common information, β, presented in Corollary 5.4.

Corollary 5.4. In the dynamic system, the broadcast transmission scheme has a better AoI performance than the unicast transmission scheme, if $\beta \le \beta_{\mathrm{th}}$, and the unicast transmission scheme has a better AoI performance than the broadcast transmission scheme, if $\beta > \beta_{\mathrm{th}}$. Specifically, β_{th} is given by

$$\beta_{\mathrm{th}} = \frac{\frac{1-e^{-\Lambda}+2\Lambda}{C_\Lambda} - \frac{3(1-e^{-\Lambda})}{C_{D_2}}}{\frac{3\Lambda}{C_{D_2}} - \frac{1-e^{-\Lambda}+2\Lambda}{C_\Lambda}} + \frac{\left(2\Lambda+1-e^{-\Lambda}\right)M_{\mathrm{L}}}{\left(\frac{3\Lambda}{C_{D_2}} - \frac{2\Lambda+1-e^{-\Lambda}}{C_\Lambda}\right)L_{\mathrm{co}}}. \tag{5.22}$$

Proof: The expected average AoI difference between these two schemes is approximated as

$$\Delta_{\mathrm{Diff}} = \frac{3}{2C_{D_2}}\left(\left(1-e^{-\Lambda}\right)L_{\mathrm{co}} + \Lambda\beta L_{\mathrm{co}}\right) - \left(\frac{1-e^{-\Lambda}}{2} + \Lambda\right)\left(\frac{(1+\beta)L_{\mathrm{co}}}{C_\Lambda} + M_{\mathrm{L}}\right). \tag{5.23}$$

Based on (5.23), we find that $\Delta_{\mathrm{Diff}} \le 0$ when $\beta \le \beta_{\mathrm{th}}$, and $\Delta_{\mathrm{Diff}} > 0$ when $\beta > \beta_{\mathrm{th}}$. ■

Corollary 5.4 reveals that the BRNP strategy achieves a better AoI performance when the common information has a high proportion of the required status update of each UE, but the UNINP strategy achieves a better AoI performance when the individual information has a high proportion of the required status update of each UE. Moreover, based on Corollary 5.4, we approximate the threshold β_{th} as

$$\beta_{\mathrm{th}} \approx \frac{2C_{D_2}}{3C_\Lambda - 2C_{D_2}}, \tag{5.24}$$

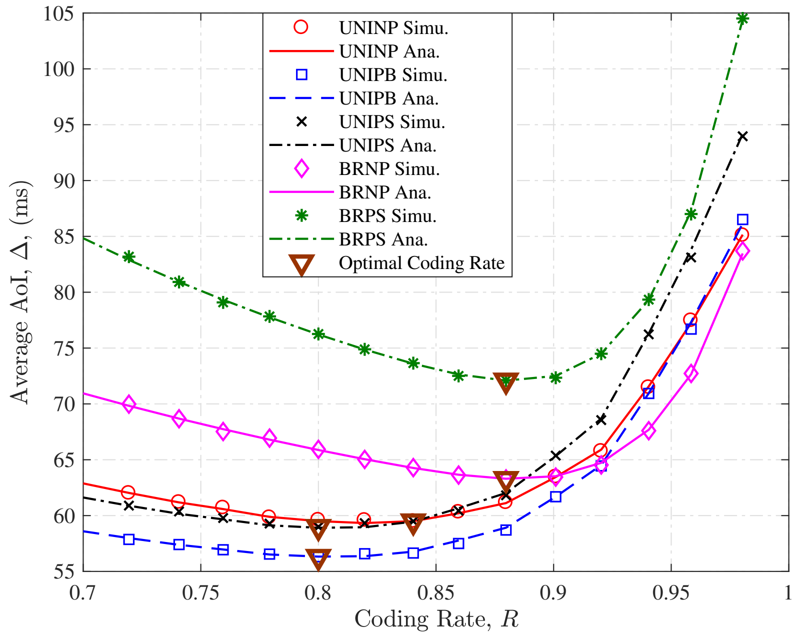
when M_{L} is small and the average number of UEs in the area, denoted by Λ, is large. We note that, in the BRNP strategy, the coding rate is based on the worst channel condition of UEs, where the increase in the coverage area of the BS leads to the significant increase in the transmission time comparing with the UNINP strategy. This indicates that the BRNP strategy achieves a better AoI performance if the BS covers a small area, where the difference in the received SNR among UEs is small, and the UNINP strategy achieves a better AoI performance if the BS covers a large area, where the difference in the received SNR among UEs is large.

5.5 Numerical Results

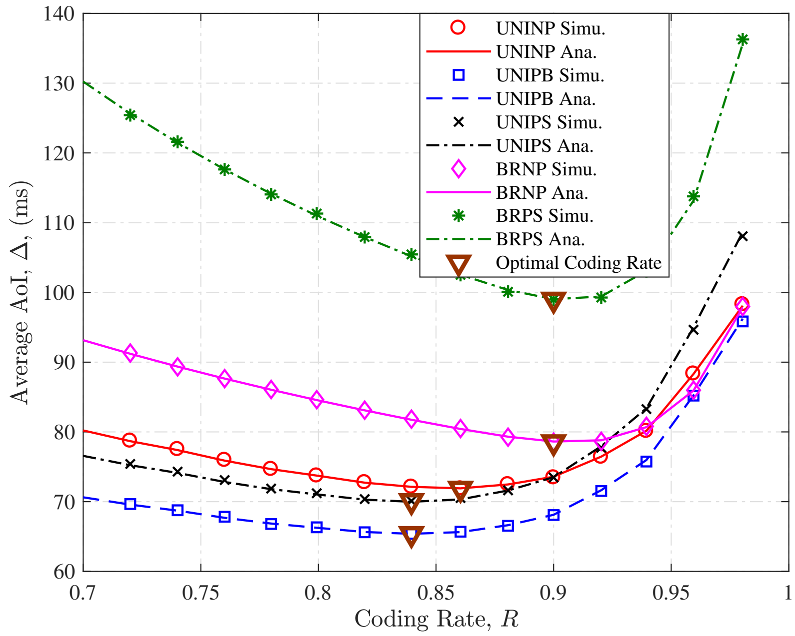
In this section, we first present numerical results and evaluate the impact of various parameters, including the coding rate, the status update generation rate, and the number of UEs on the average AoI in the homogeneous case where all UEs have the same received SNR, i.e., $\gamma_n = \gamma_h$, the same length of the required information, i.e., $L_n = L_h$, and the same coding rate, i.e., $R_n = R$, $n = 1, 2, \cdots, N$. In this case, the transmission time is calculated as $M = \alpha N L_h / R$ in the broadcast transmission scheme and the transmission time for U_n is calculated as $M_n = L_h / R$ in the unicast transmission scheme. We then examine how different transmission schemes affect the AoI performance in the considered remote control system and dynamic system. Specifically, we set the unit time for each c.u. $T_u = 0.072$ millisecond (ms) [116].

Fig. 5.3 plots the average AoI of the considered system versus the coding rate, R. We first observe that the analytical average AoI precisely matches the simulation results, which demonstrates the correctness of our analytical results in Theorem 5.1, Theorem 5.2, Theorem 5.3, and Corollary 5.2. We then observe that for all strategies, the average AoI first decreases and then increases when R increases. This observation is due to the fact that the increase in R has a two-fold effect on the average AoI. When R is small, its increase leads to a shorter packet blocklength, which decreases the average AoI of the system. When R exceeds a certain threshold, its increase leads to the significant increase in the block error rate, thereby degrading the AoI performance. We further observe that the optimal coding rate, which minimizes the average AoI, increases when the number of bits in a status update increases in all strategies. This is because that the increase in the number of bits in a status update leads to the increase in transmission time and the decrease in the block error rate. When the number of bits in a status update is low, the decrease in the block error rate dominantly and positively affects the average AoI. When the number of bits in a status update is high, the increase in transmission time dominates the average AoI, thereby degrading the AoI performance. Additionally, we observe that the BRNP strategy achieves a lower average AoI than the BRPS strategy, and the UNIPB strategy achieves the lowest average AoI comparing with the UNINP and UNIPS strategies. This observation reveals that the non-preemption strategy should be adopted in the broadcast transmission scheme and the preemption in buffer strategy should be adopted in the unicast transmission scheme.

Fig. 5.4 plots the average AoI of the considered system versus the status update generation rate, λ. We first observe that the average AoI decreases monotonically when λ increases for the UNINP, UNIPB, and BRNP strategies. This is because that the increase in λ leads to the decrease in the waiting time after successful transmission, which decreases the average AoI of the system. We then observe that the average AoI first decreases and then increases when λ increases for the UNIPS and BRPS strategies. This observation is due to the fact that the increase in λ has a two-fold effect on the average AoI. In particular, the increase



(a)



(b)

Figure 5.3: The average AoI versus the coding rate, R, with $N = 3$, $\lambda = 0.002$, $\gamma_h = 3$, $M_\mathrm{L} = 10$, and $\alpha = 1$, for (a) $L_h = 100$ and (b) $L_h = 150$.

in λ decreases the waiting time after successful transmission but increases the probability that a status update is preempted. When λ is small, the decrease in the waiting time after successful transmission dominantly and positively affects the average AoI, which improves the AoI performance. When λ exceeds a certain threshold, the probability that a status update is preempted increases significantly, thereby degrading the AoI performance.

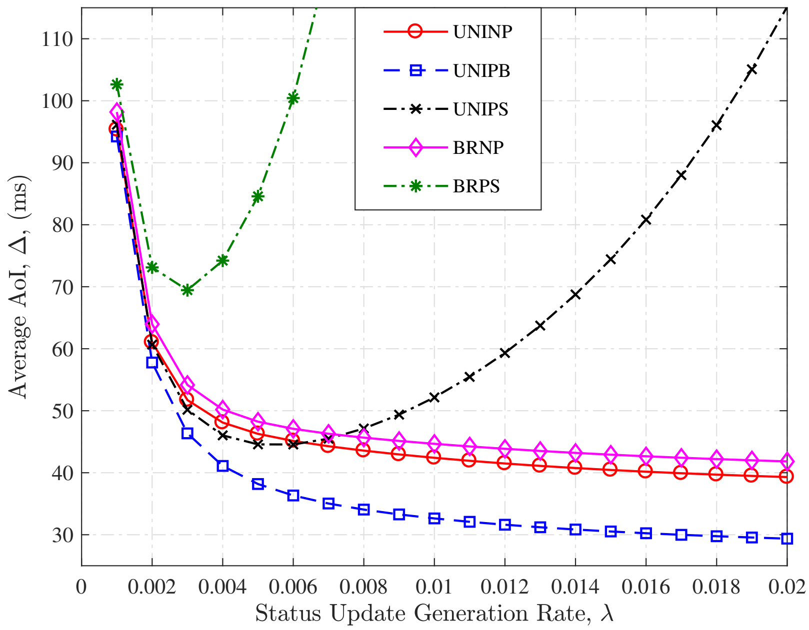
Fig. 5.5 plots the average AoI of the considered system versus the number of sensors, N. We first observe that the average AoI achieved by all strategies increases monotonically when N increases. This is because that the increase in N results in the longer transmission time of status updates for all UEs, which increases the average AoI of the system. We then observe that when N increases, the average AoI increases gently for the UNIPS strategy, but it increases dramatically for the BRPS strategy. In the BRPS strategy, the increase in N leads to the increase in the probability that a status update is preempted, which increases the average AoI. Differently, in the UNIPS strategy, the increase in N does not increase the probability that a status update is preempted for each UE, which leads to a slight increase in the average AoI. This observation also implies that the preemption in serving is more suitable in the unicast transmission scheme than in the broadcast transmission scheme.

Fig. 5.6 plots the average AoI of the remote control system versus the information ratio, α. We first observe that the analytical α_{th} precisely matches the simulation result, which demonstrates the correctness of our analytical result in Theorem 5.4. It indicates that we can adopt Theorem 5.4 to find the optimal transmission scheme to minimize the average AoI based on the information ratio. We then observe that the average AoI increases when α increases for the BRNP strategy. This is because that the increase in α leads to a larger number of bits in status updates for the transmission. Hence, the packet blocklength for transmission increases, which in turn increases the average AoI of the system. We further observe that α_{th} decreases as the number of UEs increases. This observation is due to the fact that when N is large, the UNINP strategy significantly decreases the average waiting time of the status update for each UE, comparing with the BRNP strategy, which decreases the average AoI. This observation implies that the BRNP strategy should be adopted when the number of UEs is small, while the UNINP strategy should be adopted when the number of UEs is large.

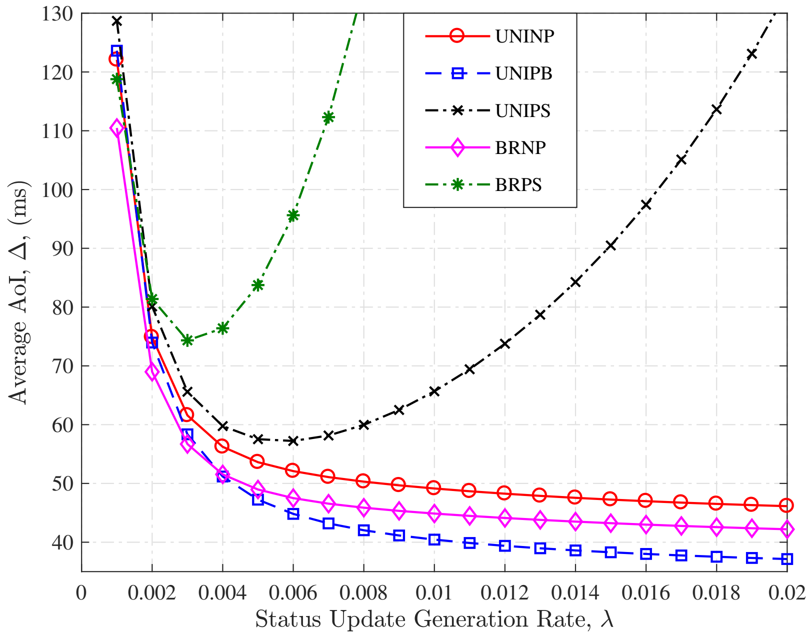
Fig. 5.7 plots the expected average AoI of the dynamic system versus the ratio between the individual information and the common information, β. We first observe that the approximation in Theorem 5.5 is close to the simulation result and there is a small gap between the approximation and the simulation result. This gap is due to the fact that the coding rate cannot achieve the channel capacity and the transmission error is inevitable in short packet communications. Due to the gap between the approximation and the simulation result, there is a gap between β_{th} in Corollary 5.4 and the exact β_{th}. We note that the gap between the expected average AoI in both schemes with β_{th} in Corollary 5.4 is small. It indicates that we can adopt Corollary 5.4 as a simple method to find the transmission scheme for minimizing the expected average AoI based on β. We further observe that the average AoI increases when β increases for both strategies. Moreover, this increase in the BRNP strategy is faster than the UNINP strategy. This is because that the increase in β leads to the increase in the length of the status update in both strategies and this increase has a more pronounced impact on the average AoI for the BRNP strategy than the UNINP strategy.

5.6 Summary

This chapter analyzed the AoI performance achieved by two transmission schemes, i.e., the broadcast transmission scheme and the unicast transmission scheme, of a multi-user system where a BS generates status updates and transmits them to multiple UEs. In both transmission schemes, we examined three packet management strategies, i.e., the non-preemption strategy, the preemption in buffer strategy, and the preemption in serving strategy. We first derived closed-form expressions for the average AoI achieved by these two transmission schemes with different packet management strategies. Such expressions allowed us to compare the AoI performance achieved by both transmission schemes in two systems, namely, the remote control system and the dynamic system. Aided by simulation results, we demonstrated the accuracy of our analysis. Moreover, we found that the unicast transmission scheme is more appropriate for a system with a large number of UEs, while the broadcast transmission scheme is more appropriate for a system with a small number of UEs and high correlation among the information for UEs.

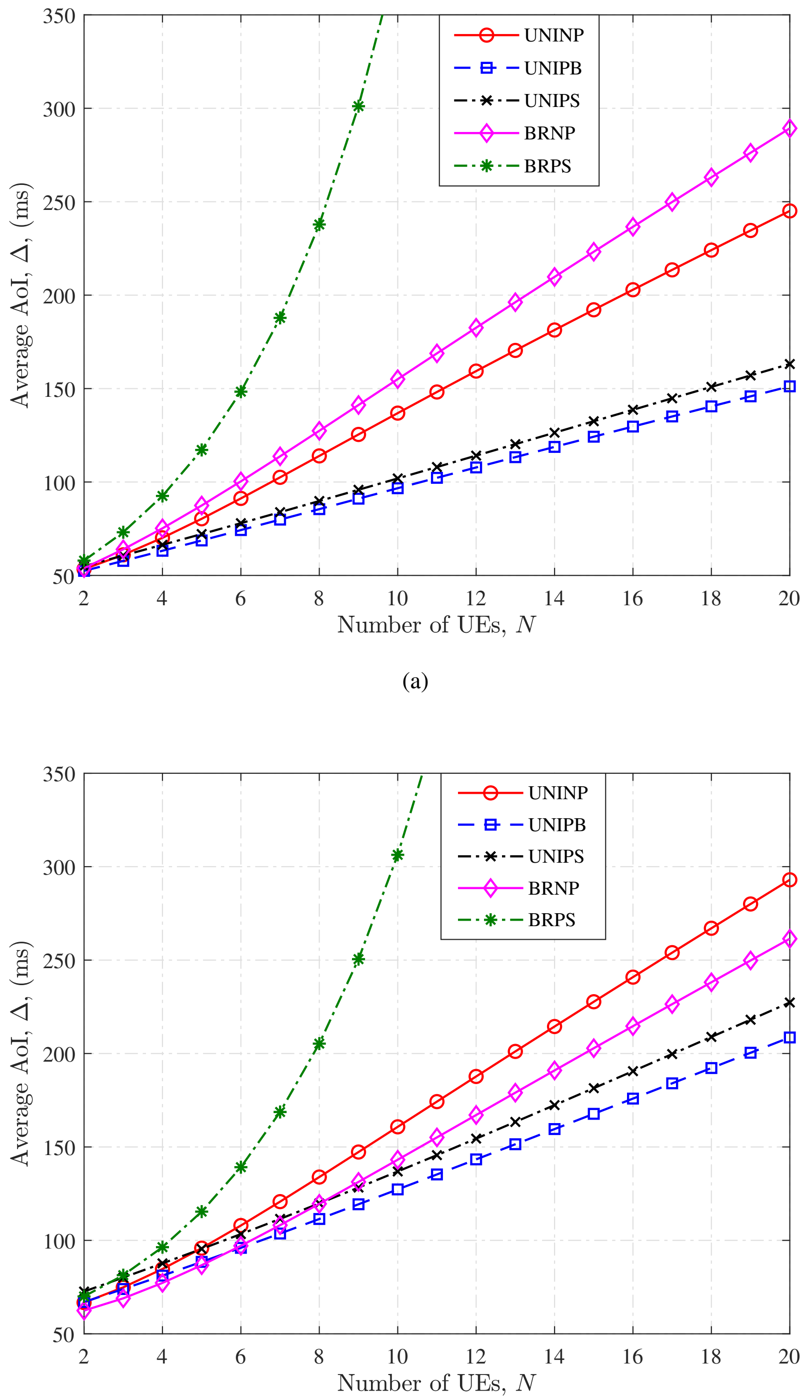


(a)



(b)

Figure 5.4: The average AoI versus the packet generation rate, λ, with $N = 3$, $L_h = 100$, $\gamma_h = 3$, $M_{\mathrm{L}} = 20$, and $\alpha = 1$ for (a) $R = 0.85$ and (b) $R = 0.95$.



(a)

(b)

Figure 5.5: The average AoI versus the number of UEs, N, with $L_h = 100$, $\lambda = 0.002$, $\gamma_h = 3$, $M_\mathrm{L} = 20$, and $\alpha = 1$ for (a) $R = 0.85$ and (b) $R = 0.95$.

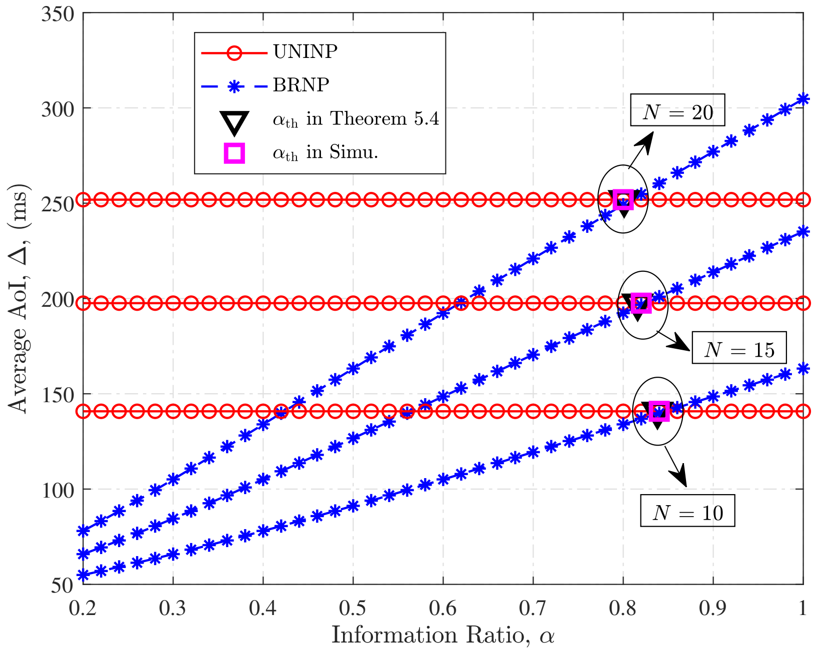


Figure 5.6: The average AoI of the remote control system versus the information ratio, α, with $L_h = 100$, $\lambda = 0.002$, $\gamma_h = 3$, $M_{\rm L} = 20$, and $R = 0.8$.

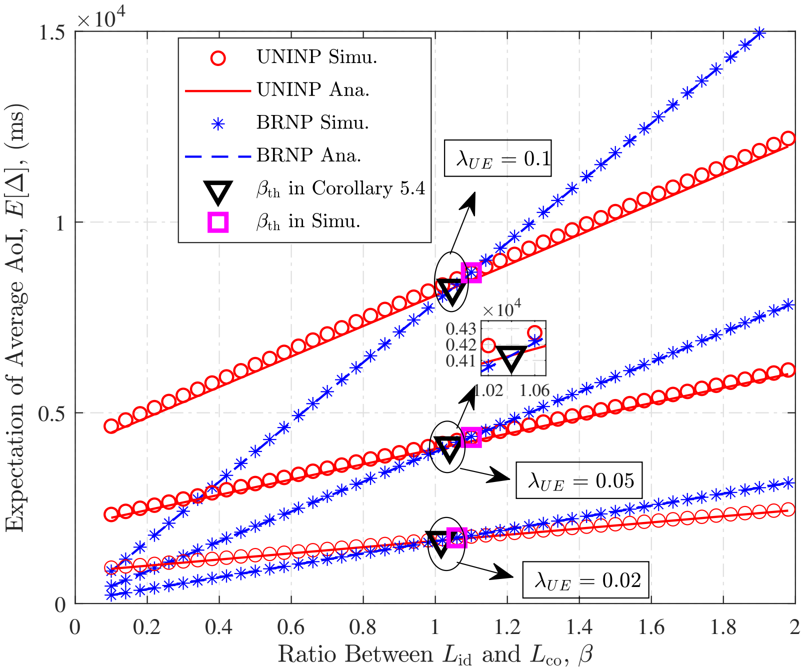


Figure 5.7: The expected average AoI of the dynamic system versus the ratio between the individual information and the common information, β, with $L_{\mathrm{co}} = 1000$, $\gamma_{D_2} = 10$, $M_{\mathrm{L}} = 10$, $D_1 = 1$, $D_2 = 20$, and $\eta = 2.2$.

Chapter 6

Conclusions

In this chapter, we first summarize the general conclusions drawn from the thesis, and then outline some future research directions arising from this thesis.

6.1 Thesis Conclusions

This thesis focused on the AoI performance and analysis in communication networks. In particular, we focused on the scheduling policy design in Chapter 2, the offloading strategy design in MEC systems in Chapter 3, the AoI performance in short packet communication systems in Chapter 4, and the AoI based transmission scheme selection in Chapter 5. The detailed contributions are given as follows.

Scheduling Policy Design for AoI Optimization: In Chapter 2, we designed a decentralized packet arrival-independent scheduling policy and a centralized packet arrival-dependent scheduling policy in uplink and downlink multi-user systems, respectively.

- We first considered an uplink multi-user system, where the packet of each UE is generated by a Bernoulli process at the beginning of each time slot and transmitted to the BS via a shared TB. We derived closed-form expressions for the average AoI for three packet management policies. Aided by simulation results, we validated our theoretical analysis and evaluated the impact of system parameters on the AoI performance. Moreover, we showed that our proposed scheduling policy outperforms the RR policy, achieving a lower average AoI. This outcome highlighted the significant benefits introduced by this proposed scheduling policy in MTC systems, paving the way for improved communication performance.

- We then considered a downlink multi-user system, where the BS generates time-sensitive packets and transmits them to UEs under an unreliable channel. To characterize the level of dissatisfaction depends on data staleness, we employed the general AoI penalty as the performance metric. By formulating the scheduling problem as an average cost CMDP problem, we derived the approximated expression for the Whittle index and proposed a

Whittle index based scheduling policy. Leveraging numerical results, we demonstrated the optimality of our proposed scheduling policy and showed that our proposed policy achieves significant AoI performance improvement when compared to existing scheduling policies.

AoI Analysis and Optimization of MEC Systems: In Chapter 3, we analyzed the average AoI and the average PAoI of a multi-user MEC system, where a BS generates and transmits computation-intensive packets to UEs. In this MEC system, we focused on three computing schemes: (i) The local computing scheme where all computational tasks are computed by the local server at the UE, (ii) The edge computing scheme where all computational tasks are computed by the edge server at the BS, and (iii) The partial computing scheme where computational tasks are partially allocated at the edge server and the rest are computed by the local server. Considering exponentially distributed transmission time and computation time and adopting the FCFS queuing policy, we derived closed-form expressions for the average AoI and PAoI for these three computing schemes. Since the expression for the average AoI is complex, we further derived simple upper and lower bounds on the average AoI, which allow us to explicitly examine the dependence of the optimal offloading decision on the MEC system parameters. In addition, we derived the optimal offloading ratio to minimize the average PAoI and found that this offloading ratio can be adopted to minimize the average AoI. Aided by simulation results, we verified our analysis and illustrated the impact of system parameters on the AoI performance and the optimal offloading strategy in the MEC system.

AoI Analysis of Short Packet Communications: In Chapter 4, we first examined the impact of different packet encoding schemes on the average AoI performance and then analyzed the non-linear AoI penalty performance of short packet communication systems.

- We first analyzed the AoI performance achieved by two encoding schemes, i.e., the joint encoding scheme and the distributed encoding scheme. In particular, we considered a point-to-point system that a transmitter transmits status updates received from multiple sensors to a receiver. In this system, these received updates can be jointly encoded into a single packet for transmission, referred to as the joint encoding scheme. Alternatively, they can be individually encoded into separate packets, referred to as the distributed encoding scheme. We derived closed-form expressions for the average AoI achieved by these two encoding schemes. Such expressions allowed us to compare the AoI performance achieved by both encoding schemes. Aided by simulation results, we demonstrated the accuracy of our analysis and found that the distributed encoding scheme is more appropriate for the system with a large number of sensors.

- We then analyzed the general average AoI penalty a point-to-point system, where a transmitter generates packets and transmits them to a receiver via an unreliable channel. We incorporated a non-linear α-β AoI penalty function and investigated the average

AoI penalty achieved by three packet management strategies, i.e., the NPNB strategy, the NPOB strategy, and the preemption strategy. Moreover, we found the conditions of the system requirement to guarantee the existence of the finite average AoI penalty. Aided by simulations, we validated the accuracy of our analytical results and established the relationship between the transmission reliability requirement and the tolerance for outdated data. This empirical verification enhances our understanding of the trade-offs between data freshness and transmission reliability in the context of our communication system.

AoI Based Transmission Scheme Selection: In Chapter 5, we analyzed the AoI performance achieved by two transmission schemes, i.e., the broadcast transmission scheme and the unicast transmission scheme, of a multi-user system where a BS generates status updates and transmits them to multiple UEs. In the broadcast transmission scheme, the status update for all UEs is jointly encoded into a packet for transmission, while in the unicast transmission scheme, the status update for each UE is encoded individually and transmitted by following the RR policy. For both transmission schemes, we examined three packet management strategies, namely the non-preemption strategy, the preemption in buffer strategy, and the preemption in serving strategy. We first derived closed-form expressions for the average AoI achieved by these two transmission schemes with different packet management strategies. Such expressions allowed us to compare the AoI performance achieved by both transmission schemes in two systems, namely, the remote control system and the dynamic system. Aided by simulation results, we verified our analysis and investigate the impact of system parameters on the average AoI. Moreover, we found that the unicast transmission scheme is more appropriate for a system with a large number of UEs, while the broadcast transmission scheme is more appropriate for a system with a small number of UEs and high correlation among the information for UEs.

The contributions presented in this thesis advance our understanding of the AoI performance in various communication networks. They also provide guidelines and methodologies to assist the communication protocol design, thereby improving the AoI performance in practical real-time applications.

6.2 Future Research Directions

The optimization of AoI performance plays a crucial role in supporting real-time applications. Given the exploratory nature of the research topic, several potential research problems naturally emerge. In this regard, a few promising research directions directly related to the thesis are briefly discussed in the following for future investigation.

- **Routing and Scheduling Policy Design:** In Chapter 2, we investigated the scheduling policy design in relatively simple multi-user systems. However, the communication network structure can often become complex, especially in large-scale deployments. It may involve multiple source-users and destinations, as well as various communication nodes and paths connecting each source-destination pair. In practice, this triggers the consideration of a multi-hop network configuration where information data is relayed through multiple intermediate nodes before reaching its final destination. In this setup, routing and scheduling policy design plays a critical role to improve the timely information delivery. While the scheduling policy governs how resources are allocated to different users or tasks at specific time instances, the routing design focuses on determining the optimal paths for information transmission between different network nodes. The development of dynamic routing algorithms that adaptively select optimal paths based on real-time network conditions are required to be explored. These algorithms should consider factors such as network structure, link quality, congestion levels, and latency to determine the most efficient routes for information transmission [119]. By dynamically adjusting routing paths in response to changing conditions, the whole real-time system can minimize latency and improve the timeliness of information delivery. This can involve techniques such as real-time path planning, collision avoidance, and mission reconfiguration to minimize end-to-end latency and maximize information delivery. Moreover, by leveraging multiple paths simultaneously, multi-path routing techniques can enhance fault tolerance and reduce the impact of link failures or congestion on information delivery. However, due to the complexity of the network structure and the dynamic nature of real-time communication systems, traditional methods often struggle to efficiently solve the problem of dynamic routing. In recent years, reinforcement learning and deep reinforcement learning have been developed as efficient and accurate tools to address complex practical problems. With these learning approaches, the exploration of efficient algorithms and protocols for selecting and utilizing multiple paths, optimizing the utilization of network resources, and ensuring timely information delivery will be one of the research directions.

- **Adaptive Offloading Strategy:** In Chapter 3, we proposed the offloading strategy in a fixed multi-user MEC system to optimize the long term average AoI. In this system, we established a system model to characterize the network topology of MEC systems without assessing the influence of both system and environmental parameters on the AoI performance. However, in real-world MEC systems, the mobility of UEs directly impacts the quality of the communication channel, with varying signal strengths and different levels of interference leading to rapid changes in channel conditions. Therefore, it is imperative to incorporate these characteristics into the MEC system model to accurately analyze the AoI performance of real-world MEC systems. Moreover, the significance of information and the complexity of computation tasks may vary among UEs based on their individual application and service requirements, making it crucial to incorporate these characters into the offloading strategy design. To address these challenges and fulfill the system requirements, it is imperative to develop a highly adaptive offloading strategy that takes into account continuous channel conditions and information importance for UEs, while accommodating device capabilities and workload demands, to optimize the system performance in real-time. However, due to high complexity of the network structure and differences between users, it is difficult to utilize the traditional optimization algorithm to optimize the AoI performance. Thus, the integration of machine learning techniques for adaptive offloading strategy design to optimize AoI performance in real-world MEC systems could be a promising future direction.

- **Age-Aware Adaptive Modulation and Scheduling Policy Design:** In Chapter 4 and Chapter 5, we investigated the AoI performance in short packet communication systems. An efficient modulation and coding technique plays a crucial role in short packet communication systems to ensure timely information delivery. Due to the propagation pathloss and fading between the transmitter and receiver, coding techniques are used to mitigate the effects of channel impairments. In short packet communications, higher order modulation schemes can achieve higher data transmission rates, while being more susceptible to errors caused by pathloss and fading. Therefore, modulation and coding technologies need to strike a balance between achieving higher data rates and providing sufficient error correction capability to overcome the pathloss and fading effects. Moreover, modulation and coding technologies design in wireless communication systems faces a significant challenge due to the mobility of users, which leads to variations in the wireless channel over time. To address these dynamic channel conditions, the adaptive modulation and coding (AMC) strategy was proposed in [120, 121, 122], which makes real-time decisions on the modulation and coding schemes based on the channel conditions to maximize the spectral efficiency, while maintaining a low block error rate. Although the AoI-aware scheduling policies were well investigated in [29, 31, 34, 36, 38], these studies typically assumed a fixed transmission time with a fixed blocklength of transmitted packets. However, when employing AMC, the transmission time may vary due to the dynamic modulation and coding schemes, which needs to be taken into account when designing scheduling policies. The integration of AMC into AoI-aware scheduling design involves joint decision making on scheduling and modulation to optimize the AoI performance of the system. In particular, the scheduler requires to consider various factors, including the AoI of different users, information bits of different packets, and the available channel conditions, and then makes decisions on user scheduling and the appropriate modulation and coding scheme for transmission. Based on these factors, the scheduler makes critical decisions regarding user scheduling and selects the appropriate modulation and coding scheme for each transmission to enhance the overall AoI performance of the system. Therefore, the joint AMC and scheduling policy design is worth being further explored, to guarantee timely information delivery in real-time applications.

Appendix A

A.1 Proof of Theorem 3.1

The proof of Theorem 3.1 relies on the following properties of exponential random variables, Poisson processes, and the $M/M/1$ queue.

Lemma A.1. Let us consider that X_1 and X_2 are independent exponential random variables with $\mathbb{E}[X_i] = 1/\gamma_i$. Given $X_1 < X_2$, the conditional PDF of variable $V = X_2 - X_1$ is given by

$$f_{V|X_1<X_2}(v|X_1 < X_2) = \gamma_2 \exp(-\gamma_2 v),\ v \geq 0. \tag{A.1}$$

Lemma A.2. Given a Poisson process $K(t)$ with the rate λ and an exponential random variable X with $\mathbb{E}[X] = 1/\gamma$, the number of arrivals $K(X)$ in the interval $[0,X]$ has the geometric PMF, which is given by

$$\Pr(K(X) = k) = (1-\alpha)\alpha^k,\ k \geq 0, \tag{A.2}$$

where $\alpha = \lambda/(\gamma+\lambda)$.

We first derive the average AoI of U_n. Based on (3.20), the average AoI of U_n in the partial computing scheme is calculated as

$$\Delta_{n,p} = \frac{1}{\lambda_n} + \frac{1}{\mu_B'} + \frac{1}{\mu_D} + \frac{1}{\mu_n'} + \lambda_n \left(\mathbb{E}[Y_j W_{j,B}] + \mathbb{E}[Y_j W_{j,D}] + \mathbb{E}[Y_j W_{j,U}]\right). \tag{A.3}$$

To obtain $\Delta_{n,p}$, we need to derive $\mathbb{E}[Y_j W_{j,B}]$, $\mathbb{E}[Y_j W_{j,D}]$, and $\mathbb{E}[Y_j W_{j,U}]$ in (A.44). We first derive $\mathbb{E}[Y_j W_{j,B}]$ as

$$\mathbb{E}[Y_j W_{j,B}] = \mathbb{E}[Y_j W_{j,B}|B_{j,B}]\Pr(B_{j,B}) + \mathbb{E}[Y_j W_{j,B}|L_{j,B}]\Pr(L_{j,B}), \tag{A.4}$$

where $B_{j,B}$ denotes the event that $P_{n,j}$ is generated before $P_{n,j-1}$ arrives at the transmission queue and $L_{j,B}$ denotes the event that $P_{n,j}$ is generated after $P_{n,j-1}$ arrives at the transmission queue. Here, $B_{j,B}$ and $L_{j,B}$ are two complementary events such that $\Pr(B_{j,B}) + \Pr(L_{j,B}) = 1$.

We first calculate $\mathbb{E}[Y_j W_{j,B}|B_{j,B}]\Pr(B_{j,B})$ in (A.4). When $B_{j,B}$ happens, $W_{j,B}$ is calculated as

$$W_{j,B} = X_{j-1,B} - Y_j + \sum_{\kappa=1}^{K_{j,y}} S_{j,\kappa,B}, \tag{A.5}$$

where $K_{j,y}$ is the number of packets generated for other UEs during Y_j and $S_{j,\kappa,B}$ is the computation time of the κth packet among $K_{j,y}$ packets computed by the edge server. As $X_{j-1,B}$ and Y_j are independent exponential random variables, according to Lemma A.1, we obtain

$$f_{V_j|B_{j,B}}(v|B_{j,B}) = (\mu_B' - \lambda)\exp(-(\mu_B' - \lambda)v), \tag{A.6}$$

where $V_j = X_{j-1,B} - Y_j$. In addition, as the packet of each UE is generated according to a Poisson process, we obtain the conditional PMF of the number of packets generated during $Y_j = y$ as

$$\Pr(K_{j,y} = k|Y_j = y) = \frac{(\lambda_{-n}y)^k \exp(-\lambda_{-n}y)}{k!}. \tag{A.7}$$

Combining (A.5), (A.6), with (A.7), we obtain

$$\mathbb{E}[W_{j,B}|Y_j = y, B_{j,B}] = \frac{1}{\mu_B' - \lambda} + \frac{\lambda_{-n}y}{\mu_B'}. \tag{A.8}$$

According to (A.8), we calculate $\mathbb{E}[Y_j W_{j,B}|B_{j,B}]\Pr(B_{j,B})$ as

$$\begin{aligned} \mathbb{E}[Y_j W_{j,B}|B_{j,B}]\Pr(B_{j,B}) &= \int_0^\infty y f_{Y_j}(y)\Pr(B_{j,B}|Y_j{=}y)\mathbb{E}[W_{j,B}|Y_j = y, B_{j,B}]\mathrm{d}y \\ &= \frac{\lambda_n}{(\mu_B' - \lambda_{-n})^2}\left(\frac{2\lambda_{-n}}{\mu_B'(\mu_B' - \lambda_{-n})} + \frac{1}{\mu_B' - \lambda}\right). \end{aligned} \tag{A.9}$$

We then calculate $\mathbb{E}[Y_j W_{j,B}|L_{j,B}]\Pr(L_{j,B})$. When $L_{j,B}$ happens, $W_{j,B}$ is calculated as

$$W_{j,B} = \sum_{\kappa=1}^{K_{j,e}} S_{j,\kappa,B}, \tag{A.10}$$

where $K_{j,e}$ is the number of packets in the computation queue at the edge server when $P_{n,j}$ is generated. Based on Lemma A.2, we obtain that when $P_{n,j}$ is generated, the PMF of the number of packets in the computation queue at the edge server is given by

$$\Pr(K_{j,e} = k) = \left(\frac{\lambda_{-n}}{\mu_B'}\right)^k \left(1 - \frac{\lambda_{-n}}{\mu_B'}\right). \tag{A.11}$$

Combining (A.10) with (A.11), we obtain

$$\mathbb{E}[W_{j,B}|Y_j = y, L_{j,B}] = \frac{\mathbb{E}[K_{j,e}]}{\mu'_B} = \frac{\lambda_{-n}}{\mu'_B(\mu'_B - \lambda_{-n})}. \tag{A.12}$$

According to (A.12), we calculate the second item in (A.4) as

$$\begin{aligned}\mathbb{E}[Y_jW_{j,B}|L_{j,B}]\Pr(L_{j,B}) &= \int_0^\infty y f_{Y_j}(y)\Pr(L_{j,B}|Y_j=y)\mathbb{E}[W_{j,B}|Y_j = y, L_{j,B}]\mathrm{d}y \\ &= \frac{\lambda_{-n}}{\mu'_B(\mu'_B - \lambda_{-n})}\left(\frac{1}{\lambda_n} - \frac{\lambda_n}{(\mu'_B - \lambda_{-n})^2}\right). \end{aligned} \tag{A.13}$$

By substituting (A.9) and (A.13) into (A.4), we obtain $\mathbb{E}[Y_jW_{j,B}]$.

Next, we derive $\mathbb{E}[Y_jW_{j,D}]$ in (A.44). We denote $Y_{j,B}$ as the time interval when $P_{n,j-1}$ and $P_{n,j}$ arrive at the transmission queue, i.e., $Y_{j,B} = t_{j,B} - t_{j-1,B}$. We then denote $B_{j,D}$ as the event that $P_{n,j}$ arrives at the transmission queue before $P_{n,j-1}$ arrives at the computation queue at the local server at U_n and $L_{j,D}$ as the event that $P_{n,j}$ arrives at the transmission queue after $P_{n,j-1}$ arrives at the computation queue at the local server at U_n. Based on these two complementary events, we calculate $\mathbb{E}[Y_jW_{j,D}]$ as

$$\mathbb{E}[Y_jW_{j,D}] = \Phi_{p,B_{j,D}} + \Phi_{p,L_{j,D}}, \tag{A.14}$$

where $\Phi_{p,B_{j,D}} = \mathbb{E}[Y_jW_{j,D}|B_{j,D}]\Pr(B_{j,D})$ and $\Phi_{p,L_{j,D}} = \mathbb{E}[Y_jW_{j,D}|L_{j,D}]\Pr(L_{j,D})$.

Here, we first calculate $\Phi_{p,B_{j,D}}$. When $B_{j,D}$ happens, $W_{j,D}$ depends on both the residual queuing delay of $P_{n,j-1}$ in the transmission queue and the transmission time of the packets generated during Y_j. Thus, we calculate $\Phi_{p,B_{j,D}}$ as

$$\Phi_{p,B_{j,D}} = \mathbb{E}[Y_j(X_{j-1,D} - Y_{j,B})|B_{j,D}]\Pr(B_{j,D}) + \mathbb{E}\left[Y_j\sum_{\kappa=1}^{K_{j,y}} S_{j,\kappa,D}\middle|B_{j,D}\right]\Pr(B_{j,D}), \tag{A.15}$$

where $S_{j,\kappa,D}$ denotes the transmission time of the κth packet among $K_{j,y}$ packets. Based on Lemma A.1, we obtain

$$\mathbb{E}[X_{j-1,D} - Y_{j,B}|B_{j,D}] = \frac{1}{\mu_D - \lambda}. \tag{A.16}$$

Since $\Pr(B_{j,D})$ is calculated as

$$\Pr(B_{j,D}) = \int_0^\infty\int_0^\infty f_{Y_j,Y_{j,B}}(y,y')\exp(-(\mu_D - \lambda)y')\mathrm{d}y'\mathrm{d}y, \tag{A.17}$$

we need to derive the conditional PDF $f_{Y_{j,B}|Y_j}(y'|y)$ to obtain $\Pr(B_{j,D})$. Here, $Y_{j,B}$ is calculated as

$$Y_{j,B} = X_{j,B} + Y_j - X_{j-1,B}, \tag{A.18}$$

where the PDF of $X_{j-1,B}$ is given as

$$f_{X_{j-1,B}}(x_1) = (\mu'_B - \lambda)\exp(-(\mu'_B - \lambda)x_1). \tag{A.19}$$

We consider two complementary events, $B_{j,B}$ and $L_{j,B}$, and derive $f_{Y_{j,B}|Y_j}(y'|y)$ as

$$f_{Y_{j,B}|Y_j}(y'|y) = p_{Y_{j,B},B_{j,B}|Y_j} + p_{Y_{j,B},L_{j,B}|Y_j}, \tag{A.20}$$

where $p_{Y_{j,B},B_{j,B}|Y_j}$ and $p_{Y_{j,B},L_{j,B}|Y_j}$ are given by

$$p_{Y_{j,B},B_{j,B}|Y_j} = f_{Y_{j,B}|Y_j,B_{j,B}}(y'|y, B_{j,B})\Pr(B_{j,B}) \tag{A.21}$$

and

$$p_{Y_{j,B},L_{j,B}|Y_j} = f_{Y_{j,B}|Y_j,L_{j,B}}(y'|y, L_{j,B})\Pr(L_{j,B}), \tag{A.22}$$

respectively.

When $B_{j,B}$ happens, $Y_{j,B}$ depends on the computation time of the packets generated during Y_j at the edge server. We consider that there are $K_{j,y} = k$ packets generated during Y_j, where the PMF of k packets generated during Y_j is given by (A.7). As the computation time of each packet follows the independent and identical exponential distribution, the total time consumed to compute these k packets and $P_{n,j}$ follows a Gamma distribution, whose PDF is given by

$$f_{Y_{j,B}|K_{j,y}}(y'|k) = \frac{y'^k \mu_B'^{k+1}\exp(-\mu'_B y')}{k!}. \tag{A.23}$$

Combining (A.7), (A.19), with (A.23), we calculate $p_{Y_{j,B},B_{j,B}|Y_j}$ as

$$\begin{aligned} p_{Y_{j,B},B_{j,B}|Y_j} &= \int_y^\infty f_{X_{j-1,D}}(x_1) \sum_{k=0}^{\infty} \Pr(K_{j,y} = k|Y_j = y) f_{Y_{j,B}|K_{j,y}}(y'|k)\mathrm{d}x_1 \\ &= \mu'_B \exp(-(\mu'_B - \lambda_n)y - \mu'_B y') I_0\left(2\sqrt{\lambda_{-n}\mu'_B y y'}\right), \end{aligned} \tag{A.24}$$

where $I_0(\cdot)$ is the modified first-kind Bessel function of the zeroth order.

When $L_{j,B}$ happens, $Y_{j,B}$ depends on the time interval $Y_j - X_{j-1,B}$ and $X_{j,B}$. In particular, $X_{j,B}$ depends on the number of the packets in the computation queue at the edge server when $P_{n,j}$ is generated. We consider that when $P_{n,j}$ is generated, there are $K_{j,e} = k$ packets in the computation queue at the edge server. As the computation time of each packet follows the independent and identical exponential distribution, the total time consumed to compute these k packets and $P_{n,j}$ follows a Gamma distribution, whose PDF is given by

$$f_{X_{j,B}|K_{j,e}}(x_2|k) = \frac{x_2^k \mu_B'^{k+1} \exp(-\mu_B' x_2)}{k!}. \tag{A.25}$$

Combining (A.11), (A.19), with (A.25), we calculate $p_{Y_{j,B},L_{j,B}|Y_j}$ as

$$\begin{aligned} p_{Y_{j,B},L_{j,B}|Y_j} = &\frac{(\mu_B'-\lambda)(\mu_B'-\lambda_{-n})}{(2\mu_B'-\lambda-\lambda_{-n})}\big(\exp(-(\mu_B'-\lambda)(y-y')) \\ &-\exp(-(\mu_B'-\lambda)y-(\mu_B'-\lambda_{-n})y')\big), \end{aligned} \tag{A.26}$$

for $y' < y$, and

$$\begin{aligned} p_{Y_{j,B},L_{j,B}|Y_j} = &\frac{(\mu_B'-\lambda)(\mu_B'-\lambda_{-n})}{(2\mu_B'-\lambda-\lambda_{-n})}\big(\exp(-(\mu_B'-\lambda_{-n})(y'-y)) \\ &-\exp(-(\mu_B'-\lambda)y-(\mu_B'-\lambda_{-n})y')\big), \end{aligned} \tag{A.27}$$

for $y' \geq y$. Combining (A.24), (A.26), with (A.27), we obtain $f_{Y_{j,B}|Y_j}(y'|y)$ in (A.20). By substituting (A.20) into (A.17) and combining (A.17) with (A.16), we obtain the first item in (A.15) as

$$\begin{aligned} \mathbb{E}[Y_j(X_{j-1,D}-Y_{j,B})|B_{j,D}]\Pr(B_{j,D}) = &\frac{\lambda_n(\mu_B'+\mu_D-\lambda)}{\mu_B'(\mu_D-\lambda)(\mu_B'+\mu_D-\lambda-\lambda_{-n})^2} \\ &+\frac{\lambda_n(\mu_B'-\lambda)(\mu_B'-\lambda_{-n})\left(\frac{1}{(\mu_D-\lambda_{-n})^2}-\frac{1}{(\mu_B'-\lambda_{-n})^2}\right)}{(\mu_B'-\mu_D)(\mu_D-\lambda)(\mu_D+\mu_B'-\lambda-\lambda_{-n})}. \end{aligned} \tag{A.28}$$

We further calculate the second item in (A.15). As the PMF of $K_{j,y}$ is given by (A.7), we obtain

$$\begin{aligned} \mathbb{E}\left[Y_j\sum_{\kappa=1}^{K_{j,y}} S_{j,\kappa,D}\middle|B_{j,D}\right]\Pr(B_{j,D}) = &\frac{2\lambda_n\lambda_{-n}(\mu_D+\mu_B'-\lambda)}{\mu_B'\mu_D(\mu_D+\mu_B'-\lambda-\lambda_{-n})^3} \\ &+\frac{\lambda_n\lambda_{-n}(\mu_B'-\lambda)(\mu_B'-\lambda_{-n})\left(\frac{2}{(\mu_D-\lambda_{-n})^3}-\frac{2}{(\mu_B'-\lambda_{-n})^3}\right)}{\mu_D(\mu_B'-\mu_D)(\mu_B'+\mu_D-\lambda-\lambda_{-n})}. \end{aligned} \tag{A.29}$$

By substituting (A.29) and (A.28) into (A.15), we obtain $\Phi_{p,B_{j,D}}$ given by (3.26).

We then calculate $\Phi_{p,L_{j,D}}$ in (A.14). When $L_{j,D}$ happens, $W_{j,D}$ depends on the transmission time of the packets in the transmission queue when $P_{n,j}$ arrives at the transmission queue. Based on Lemma A.2, we obtain

$$\mathbb{E}[W_{j,D}|Y_j=y,L_{j,D}] = \mathbb{E}\left[\sum_{\kappa=1}^{K_{j,T}} S_{j,\kappa,D}\middle|Y_j=y,L_{j,D}\right] = \frac{\lambda_{-n}}{\mu_D(\mu_D-\lambda_{-n})}, \tag{A.30}$$

where $K_{j,T}$ is the number of packets in the transmission queue when $P_{n,j}$ arrives at the transmission queue. Thus, we calculate $\Phi_{p,L_{j,D}}$ as

$$\Phi_{p,L_{j,D}} = \int_0^\infty \int_0^\infty \mathbb{E}[W_{j,D}|Y_j = y, L_{j,D}] y f_{Y_j}(y) \times f_{Y_{j,B}|Y_j}(y'|y) \Pr(L_{j,D}|Y_j = y, Y_{j,B} = y') \mathrm{d}y' \mathrm{d}y, \tag{A.31}$$

and then obtain $\Phi_{p,L_{j,D}}$ given by (3.27). By substituting (3.26) and (3.27) into (A.14), we obtain $\mathbb{E}[Y_j W_{j,B}]$.

Finally, we derive $\mathbb{E}[Y_j W_{j,U}]$ in (A.44). Here, we denote $Y_{j,D}$ as the time interval when $P_{n,j-1}$ and $P_{n,j}$ arrive at the computation queue at the local server at U_n, i.e., $Y_{j,D} = t_{j,D} - t_{j-1,D}$. Since $\mathbb{E}[W_{j,U} Y_j]$ is calculated as

$$\mathbb{E}[W_{j,U} Y_j] = \int_0^\infty \int_0^\infty y f_{Y_j}(y) f_{Y_{j,D}|Y_j}(y''|y) \mathbb{E}[W_{j,U}|Y_j = y, Y_{j,D} = y''] \mathrm{d}y'' \mathrm{d}y, \tag{A.32}$$

we need to derive $f_{Y_{j,D}|Y_j}(y''|y)$ and $\mathbb{E}[W_{j,U}|Y_j = y, Y_{j,D} = y'']$ to obtain $\mathbb{E}[Y_j W_{j,U}]$. Here, $W_{j,U}$ only depends on the computation time of $P_{n,j-1}$ at the local server. Then, we obtain

$$\mathbb{E}[W_{j,U}|Y_j = y, Y_{j,D} = y''] = \frac{\exp(-(\mu_n' - \lambda_n) y'')}{\mu_n' - \lambda_n}. \tag{A.33}$$

Since $B_{j,D}$ and $L_{j,D}$ are two complementary events, we rewrite $f_{Y_{j,D}|Y_j}(y''|y)$ as

$$f_{Y_{j,D}|Y_j}(y''|y) = p_{Y_{j,D},B_{j,D}|Y_j} + p_{Y_{j,D},L_{j,D}|Y_j}, \tag{A.34}$$

where $p_{Y_{j,D},B_{j,D}|Y_j}$ and $p_{Y_{j,D},L_{j,D}|Y_j}$ are given by

$$p_{Y_{j,D},B_{j,D}|Y_j} = f_{Y_{j,D}|Y_j,B_{j,D}}(y''|y, B_{j,D}) \Pr(B_{j,D}|Y_j = y) \tag{A.35}$$

and

$$p_{Y_{j,D},L_{j,D}|Y_j} = f_{Y_{j,D}|Y_j,L_{j,D}}(y''|y, L_{j,D}) \Pr(L_{j,D}|Y_j = y), \tag{A.36}$$

respectively. Here, we calculate $p_{Y_{j,D},B_{j,D}|Y_j}$ as

$$p_{Y_{j,D},B_{j,D}|Y_j} = f_{Y_{j,D}|Y_j,B_{j,D}}(y''|y, B_{j,D}) \int_0^\infty f_{Y_{j,B}|Y_j}(y'|y) \exp(-(\mu_D - \lambda) y') \mathrm{d}y'. \tag{A.37}$$

When $B_{j,D}$ happens, the time interval $Y_{j,D}$ depends on the transmission time of the packets generated during Y_j and the transmission time of $P_{n,j}$. Note that, the PMF of the number of packets generated during Y_j for other UEs is given by (A.7) and the PDF of the total transmission time of $k+1$ packets is given by (A.23). Thus, we obtain

$$f_{Y_{j,D}|Y_j,B_{j,D}}(y''|y,B_{j,D}) = \sum_{k=0}^{\infty} \Pr(K_{j,y}=k) f_{Y_{j,D}|K_{j,y}}(y''|k)$$
$$= \mu_D \exp(-\lambda_{-n} y - \mu_D y'') I_0(2\sqrt{\lambda_{-n}\mu_D y y''}). \tag{A.38}$$

By substituting $f_{Y_{j,B}|Y_j}(y'|y)$ and (A.38) into (A.37), we obtain $p_{Y_{j,D},B_{j,D}|Y_j}$ in (A.34). We next calculate $p_{Y_{j,D},L_{j,D}|Y_j}$ in (A.34) as

$$f_{Y_{j,D}|Y_j,L_{j,D}}(y''|y,L_{j,D}) \Pr(L_{j,D}|Y_j=y) = \int_0^{\infty} f_{Y_{j,B}|Y_j}(y'|y) p_{Y_{j,D},L_{j,D}|Y_j,Y_{j,B}} \mathrm{d}y', \tag{A.39}$$

where

$$p_{Y_{j,D},L_{j,D}|Y_j,Y_{j,B}} = f_{Y_{j,D}|Y_j,Y_{j,B},L_{j,D}}(y''|y,y',L_{j,D}) \Pr(L_{j,D}|Y_j=y,Y_{j,B}=y'). \tag{A.40}$$

Since $p_{Y_{j,D},L_{j,D}|Y_j,Y_{j,B}}$ depends on the transmission time of $P_{n,j}$ and the transmission time of packets in the transmission queue when $P_{n,j}$ arrives at the transmission queue, we obtain

$$p_{Y_{j,D},L_{j,D}|Y_j,Y_{j,B}} = \frac{(\mu_D-\lambda)(\mu_D-\lambda_{-n})}{(2\mu_D-\lambda-\lambda_{-n})} \big(\exp(-(\mu_D-\lambda)(y'-y''))$$
$$- \exp(-(\mu_D-\lambda)y' - (\mu_D-\lambda_{-n})y'')\big), \tag{A.41}$$

for $y''<y'$, and

$$p_{Y_{j,D},L_{j,D}|Y_j,Y_{j,B}} = \frac{(\mu_D-\lambda)(\mu_D-\lambda_{-n})}{(2\mu_D-\lambda-\lambda_{-n})} \big(\exp(-(\mu_D-\lambda_{-n})(y''-y'))$$
$$- \exp(-(\mu_D-\lambda)y' - (\mu_D-\lambda_{-n})y'')\big), \tag{A.42}$$

for $y'' \geq y'$. By substituting $f_{Y_{j,B}|Y_j}(y'|y)$, (A.41), and (A.42) into (A.39), we obtain $p_{Y_{j,D},B_{j,D}|Y_j}$ in (A.34). In addition, by substituting (A.33) and (A.34) into (A.32), we obtain

$$\mathbb{E}[W_{j,U}Y_j] = \Phi_{p,B_{j,U}} + \Phi_{p,L_{j,U}}, \tag{A.43}$$

where $\Phi_{p,B_{j,U}}$ is given by (3.28) and $\Phi_{p,L_{j,U}}$ is given by (3.29). Finally, we substitute (A.4), (A.14), and (A.43) into (A.44) and obtain the average AoI of U_n, which is given in (3.24).

We then drive the average PAoI. Based on (3.21), the average AoI of U_n in the partial computing scheme is calculated as

$$\Omega_{n,p} = \frac{1}{\lambda_n} + \frac{1}{\mu'_B} + \frac{1}{\mu_D} + \frac{1}{\mu'_n} + \mathbb{E}[W_{j,B}] + \mathbb{E}[W_{j,D}] + \mathbb{E}[W_{j,U}]. \tag{A.44}$$

According to Lemma A.2, the PMF of the number of packets in the computation queue at the edge server when $P_{n,j}$ is generated is given by

$$\Pr(K_{j,e}=k)=\left(\frac{\lambda}{\mu_B'}\right)^k\left(1-\frac{\lambda}{\mu_B'}\right). \tag{A.45}$$

Hence, we obtain

$$\mathbb{E}[W_{j,B}]=\mathbb{E}[K_{j,e}]\mathbb{E}[S_D]=\frac{\lambda}{\mu_B'(\mu_B'-\lambda)}. \tag{A.46}$$

Similarly, we calculate $\mathbb{E}[W_{j,D}]$ and $\mathbb{E}[W_{j,U}]$ by

$$\mathbb{E}[W_{j,D}]=\frac{\lambda}{\mu_D(\mu_D-\lambda)}, \tag{A.47}$$

and

$$\mathbb{E}[W_{j,U}]=\frac{\lambda_n}{\mu_n'(\mu_n'-\lambda_n)}. \tag{A.48}$$

By substituting (A.46), (A.47), and (A.48) into (3.21), we obtain the average PAoI of U_n, which is given in (3.25).

A.2 Proof of Theorem 3.2

We first derive the upper bound on the average AoI. Based on (3.20), we find that $W_{j,B}$, $W_{j,D}$, and $W_{j,U}$ are negatively correlated with Y_j, respectively, i.e., $\mathbb{E}[W_{j,B}Y_j]<\mathbb{E}[W_{j,B}]\mathbb{E}[Y_j]$, $\mathbb{E}[W_{j,D}Y_j]<\mathbb{E}[W_{j,D}]\mathbb{E}[Y_j]$, and $\mathbb{E}[W_{j,U}Y_j]<\mathbb{E}[W_{j,U}]\mathbb{E}[Y_j]$. Hence, we obtain that

$$\Delta_n<\frac{1}{\lambda_n}+\mathbb{E}[S_{j,B}]+\mathbb{E}[S_{j,D}]+\mathbb{E}[S_{j,U}]+\lambda_n\mathbb{E}[Y_j]\left(\mathbb{E}[W_{j,B}]+\mathbb{E}[W_{j,D}]+\mathbb{E}[W_{j,U}]\right)=\Omega_n. \tag{A.49}$$

We then derive the lower bound on the average AoI. Based on (A.4), we obtain

$$\begin{aligned}\mathbb{E}[Y_jW_{j,B}]=&\frac{\lambda_{-n}}{\lambda_n\mu_B'(\mu_B'-\lambda_{-n})}+\frac{\lambda_n\lambda_{-n}}{\mu_B'(\mu_B'-\lambda_{-n})^3}+\frac{\lambda_n}{(\mu_B'-\lambda_{-n})^2(\mu_B'-\lambda)}\\=&\frac{\lambda}{\lambda_n\mu_B'(\mu_B'-\lambda)}+\frac{\lambda_n\lambda_{-n}}{\mu_B'(\mu_B'-\lambda_{-n})^3}-\frac{1}{(\mu_B'-\lambda_{-n})^2}.\end{aligned} \tag{A.50}$$

We find that $\mathbb{E}[Y_jW_{j,D}]$ in (A.14) decreases as μ_B' increases. Hence, we obtain

$$\begin{aligned}\mathbb{E}[Y_jW_{j,D}]\geq&\,\mathbb{E}[Y_jW_{j,D}]\big|_{\mu_B'\to\infty}=\mathbb{E}[Y_{j,B}W_{j,D}]\\=&\frac{\lambda}{\lambda_n\mu_D(\mu_D-\lambda)}+\frac{\lambda_n\lambda_{-n}}{\mu_D(\mu_D-\lambda_{-n})^3}-\frac{1}{(\mu_D-\lambda_{-n})^2}.\end{aligned} \tag{A.51}$$

Similarly, we obtain the lower bound on $\mathbb{E}[Y_j W_{j,U}]$, which is given by

$$\mathbb{E}[Y_j W_{j,U}] \geq \mathbb{E}[Y_j W_{j,U}]\big|_{\mu'_B \to \infty, \mu_D \to \infty} = \mathbb{E}[Y_{j,D} W_{j,U}] = \frac{1}{\mu'_n(\mu'_n - \lambda_n)} - \frac{1}{\mu'^2_n}. \tag{A.52}$$

By substituting (A.50), (A.51), and (A.52) into (3.20), and averaging Δ_n over all UEs, we obtain the lower bound on the average AoI, which is given in (3.40).

A.3 Proof of Lemma 3.1

By calculating the first and the second derivatives of Ω w.r.t. p, we obtain

$$\frac{\partial \Omega}{\partial p} = \frac{\mu_B}{(\mu_B - p\lambda)^2} - \frac{1}{N} \sum_{n=1}^{N} \frac{\mu_n}{(\mu_n - (1-p)\lambda_n)^2} \tag{A.53}$$

and

$$\frac{\partial^2 \Omega}{\partial p^2} = \frac{2\lambda \mu_B}{(\mu_B - p\lambda)^3} + \frac{1}{N} \sum_{n=1}^{N} \frac{2(1-p)\lambda_n^2}{(\mu_n - (1-p)\lambda_n)^3}, \tag{A.54}$$

respectively. From (A.54), we see that $\frac{\partial^2 \Omega}{\partial p^2} > 0$ when $p < \min\{\frac{\mu_B}{\lambda}, 1\}$ and $p > \max\{\frac{\mu_n - \lambda_n}{\lambda_n}, 0\}$ for $n = 1, 2, \cdots, N$. This implies that there exists an optimal value of p to minimize Ω. By simple mathematical manipulation, we reach the solution to (3.44).

A.4 Proof of Theorem 3.3

According to (A.53) and (A.54), the first derivatives of Ω w.r.t. p in the homogeneous MEC system is calculated as

$$\frac{\partial \Omega}{\partial p} = \frac{\mu_B}{(\mu_B - p\lambda)^2} - \frac{\mu_h}{(\mu_h - (1-p)\lambda_h)^2}. \tag{A.55}$$

Based on (A.55), we obtain the optimal p, which is given as

$$p_{\text{opt}} = \frac{\sqrt{\mu_B \mu_n} + \lambda_n - \mu_n}{\left(1 + N\sqrt{\frac{\mu_n}{\mu_B}}\right)\lambda_n}. \tag{A.56}$$

By setting $p \in [0, 1]$ in (A.56), we obtain the optimal p_{opt} given by (3.45).

Appendix B

B.1 Proof of Theorem 5.1

To obtain $\Delta_{\text{UNINP},n}$, we need to derive $\mathbb{E}[T_j]$, $\mathbb{E}[\zeta_j]$, and $\mathbb{E}[\zeta_j^2]$ in (5.6). We first derive $\mathbb{E}[T_j]$ by calculating it as

$$\mathbb{E}[T_j] = \mathbb{E}[S_j] + \mathbb{E}[W_j] = \sum_{k=1}^{n} M'_k + \mathbb{E}[W_j], \tag{B.1}$$

where S_j is the serving time of P_j from the time that P_j is served by the BS to the time that U_n receives P_j, and W_j is the waiting time of P_j in the buffer. Specifically, W_j is the time interval from the generation time of P_j to the time that P_j is served by the BS. We note that, if P_j is generated when the BS is in the idle state, $W_j = 0$ and its probability is given by $\Pr(W_j = 0) = e^{-\lambda M_{\text{T}}}$. Otherwise, W_j follows an exponential distribution over the range $(0, M_{\text{T}})$ with the PDF given by $f_{W_j}(w) = \lambda e^{-\lambda w}$, since the LCFS queuing policy is adopted and the packet generation follows a Poisson process. Then, we derive $\mathbb{E}[W_j]$ as

$$\mathbb{E}[W_j] = \int_0^{M_{\text{T}}} w\lambda e^{-\lambda w} \mathrm{d}w = \frac{1 - e^{-\lambda M_{\text{T}}}}{\lambda} - M_{\text{T}} e^{-\lambda M_{\text{T}}}. \tag{B.2}$$

By substituting (B.2) into (B.1), we obtain $\mathbb{E}[T_j]$.

We then derive $\mathbb{E}[\zeta_j]$. Here, we denote H_j as the number of status update transmitted to U_n after P_{j-1} is received by U_n and before P_j is received by U_n, and denote $P_{j,h}$ as the hth status update transmitted to U_n after P_{j-1}, where $P_{j,H_j} = P_j$. We note that

$$\mathbb{E}[\zeta_j] = \mathbb{E}\left[\sum_{h=1}^{H_j} B_{j,h}\right] = \mathbb{E}[H_j]\mathbb{E}[B_{j,h}], \tag{B.3}$$

where $B_{j,h}$ is the time interval from the time that $P_{j,h-1}$ is transmitted to U_n to the time that $P_{j,h}$ is transmitted to U_n, since $B_{j,h}$ is independently and identically distributed for $\forall h$. Here, we note that if $h = 1$, $P_{j,0} = P_{j-1}$. We find that H_j follows a geometric distribution, where the probability mass function (PMF) of H_j is given by

$$\Pr(H_j = H) = (1-\varepsilon_n)\varepsilon_n^{H-1}. \tag{B.4}$$

According to (B.4), we obtain the expectation of H_j and H_j^2, i.e., $\mathbb{E}[H_j]$ and $\mathbb{E}[H_j^2]$, are

$$\mathbb{E}[H_j] = \frac{1}{1-\varepsilon_n}, \tag{B.5}$$

and

$$\mathbb{E}[H_j^2] = \frac{1+\varepsilon_n}{(1-\varepsilon_n)^2}, \tag{B.6}$$

respectively. In (B.3), $B_{j,h}$ is calculated by

$$B_{j,h} = V_{j,h} + M_{\mathrm{T}}, \tag{B.7}$$

where $V_{j,h}$ is the time that the BS is in the idle state during $B_{j,h}$. We note that $V_{j,h}$ exists if and only if $P_{j,h}$ is generated when the BS is in the idle state after $P_{j,h-1}$ is served. Since $V_{j,h}$ follows an exponential distribution, we obtain $\mathbb{E}[V_{j,h}]$ as

$$\mathbb{E}[V_{j,h}] = \int_{M_{\mathrm{T}}}^{\infty}(x-M_{\mathrm{T}})\lambda e^{-\lambda x}\mathrm{d}x = \frac{1}{\lambda}e^{-\lambda M_{\mathrm{T}}}. \tag{B.8}$$

By substituting (B.5), (B.7), and (B.8) into (B.3), we obtain $\mathbb{E}[\zeta_j]$ as

$$\mathbb{E}[\zeta_j] = \frac{1}{1-\varepsilon_n}\left(M_{\mathrm{T}} + \frac{e^{-\lambda M_{\mathrm{T}}}}{\lambda}\right). \tag{B.9}$$

We next derive $\mathbb{E}\left[\zeta_j^2\right]$. Due to the independent and identical distribution of B_{j,h_1} and B_{j,h_2} when $h_1 \neq h_2$, we calculate $\mathbb{E}\left[\zeta_j^2\right]$ as

$$\mathbb{E}\left[\zeta_j^2\right] = \mathbb{E}\left[\left(\sum_{h=1}^{H_j} B_{j,h}\right)^2\right] = \mathbb{E}[H_j]\mathbb{E}\left[B_{j,h}^2\right] + \mathbb{E}\left[H_j^2 - H_j\right]\mathbb{E}\left[B_{j,h}\right]^2. \tag{B.10}$$

Since $V_{j,k}$ follows an exponential distribution, we obtain $\mathbb{E}\left[V_{j,h}^2\right]$ as

$$\mathbb{E}\left[V_{j,h}^2\right] = \int_{M_{\mathrm{T}}}^{\infty}(x-M_{\mathrm{T}})^2\lambda e^{-\lambda x}\mathrm{d}x = \frac{2}{\lambda^2}e^{-\lambda M_{\mathrm{T}}}. \tag{B.11}$$

By substituting (B.5), (B.6), (B.7), and (B.11) into (B.10), we obtain $\mathbb{E}\left[\zeta_j^2\right]$ as

$$\mathbb{E}\left[\zeta_j^2\right] = \frac{\left(M_\mathrm{T} + \frac{e^{-\lambda M_\mathrm{T}}}{\lambda}\right)^2 (1+\varepsilon_n)}{(1-\varepsilon_n)^2} + \frac{2e^{-\lambda M_\mathrm{T}} - e^{-2\lambda M_\mathrm{T}}}{(1-\varepsilon_n)\lambda^2}. \tag{B.12}$$

By substituting (B.1), (B.9), and (B.12) into (5.6), we obtain the final result given in (5.8), which completes this proof.

B.2 Proof of Theorem 5.2

Based on (5.6), we need to calculate $\mathbb{E}[T_j]$, $\mathbb{E}[\zeta_j]$, and $\mathbb{E}[\zeta_j^2]$ to obtain $\Delta_{\mathrm{UNIPB},n}$. We first derive $\mathbb{E}[T_j]$ by calculating it as

$$\mathbb{E}[T_j] = \mathbb{E}[S_j] + \mathbb{E}[W_j]. \tag{B.13}$$

Here, we denote $A_{j,k}$ as the event that P_j received by U_n is generated when the BS is serving $U_{g_{k,n}}$ and $A_{j,0}$ as the event that P_j received by U_n is generated when the BS is in the idle state. Then, the probability of $A_{j,k}$ is obtained as

$$\Pr(A_{j,k}) = \begin{cases} p_k \prod_{\kappa=k+1}^{N-1} (1-p_\kappa), & \text{if } k \neq N \text{ and } k \neq 0, \\ \prod_{\kappa=1}^{N} (1-p_\kappa), & \text{if } k = 0, \\ p_N \prod_{\kappa=1}^{N-1} (1-p_\kappa), & \text{otherwise,} \end{cases} \tag{B.14}$$

where $p_k = 1 - e^{-\lambda M'_{g_{k,n}}}$ is the probability that at least one status update is generated when the BS is serving $U_{g_{k,n}}$. Hence, $\mathbb{E}[W_j]$ is calculated as

$$\mathbb{E}[W_j] = \sum_{k=1}^{N} \frac{\Pr(A_{j,k})}{p_k} \int_0^{M'_{g_{k,n}}} x\lambda e^{-\lambda x} \mathrm{d}x = \frac{1-e^{-\lambda M_\mathrm{T}}}{\lambda} - \sum_{k=1}^{N-1} M'_{g_{k,n}} e^{-\lambda \sum_{\kappa=k}^{N-1} M'_{g_{\kappa,n}}} - M'_{g_{N,n}} e^{-\lambda M_\mathrm{T}}. \tag{B.15}$$

We denote $G_{j,k}$ as the event that the first packet of P_j is transmitted to $U_{g_{k,n}}$, where the probability of $G_{j,k}$ is given by

$$\Pr(G_{j,k}) = \begin{cases} \frac{\Pr(A_{j,k-1})}{1-e^{-\lambda M_\mathrm{T}}}, & \text{if } k \neq 1, \\ \frac{\Pr(A_{j,N})}{1-e^{-\lambda M_\mathrm{T}}}, & \text{otherwise.} \end{cases} \tag{B.16}$$

Based on (B.16), we calculate $\mathbb{E}[S_j]$ by

$$\mathbb{E}[S_j] = \sum_{k=1}^{N} \Pr(G_{j,k}) \sum_{\kappa=k}^{N} M'_{g_{\kappa,n}} = \sum_{k=1}^{N} M'_{g_{k,n}} \frac{e^{-\lambda \sum_{\kappa=k}^{N-1} M'_{g_{\kappa,n}}} - e^{-\lambda M_\mathrm{T}}}{1 - e^{-\lambda M_\mathrm{T}}}. \tag{B.17}$$

By substituting (B.15) and (B.17) into (B.13), we obtain $\mathbb{E}[T_j]$ as

$$\mathbb{E}[T_j] = \frac{1-e^{-\lambda M_{\mathrm{T}}}}{\lambda} + M'_{g_{N,n}}(1-e^{-\lambda M_{\mathrm{T}}}) - \sum_{k=1}^{N} M'_{g_{k,n}} \left(\frac{e^{-\lambda M_{\mathrm{T}}}\left(1-e^{-\lambda \sum_{\kappa=k}^{N-1} M'_{g_{\kappa,n}}}\right)}{1-e^{-\lambda M_{\mathrm{T}}}} \right). \tag{B.18}$$

We then derive $\mathbb{E}[\zeta_j]$ based on (B.3) and (B.7). We denote $Z_{j,k,h}$ as the event that the last packet of $P_{j,h-1}$ is for $U_{g_{k,n}}$ and the BS does not generate any status update during the service of $P_{j,h-1}$. Then, the probability of $Z_{j,k,h}$ is given by

$$\Pr(Z_{j,k,h}) = \frac{\left(1-e^{-\lambda M'_{g_{k,n}}}\right)e^{-\lambda M_{\mathrm{T}}}}{1-e^{-\lambda M_{\mathrm{T}}}}. \tag{B.19}$$

Hence, we calculate $\mathbb{E}[V_{j,h}]$ as

$$\mathbb{E}[V_{j,h}] = \frac{1}{\lambda}\sum_{k=1}^{N} \Pr(Z_{j,k,h}) = \frac{e^{-\lambda M_{\mathrm{T}}}}{\lambda(1-e^{-\lambda M_{\mathrm{T}}})} \sum_{k=1}^{N}\left(1-e^{-\lambda M'_{g_{k,n}}}\right). \tag{B.20}$$

By substituting (B.20) into (B.7) and the result in (B.3), we obtain $\mathbb{E}[\zeta_j]$ as

$$\mathbb{E}[\zeta_j] = \frac{1}{1-\varepsilon_n}\left(M_{\mathrm{T}} + \frac{e^{-\lambda M_{\mathrm{T}}}}{\lambda(1-e^{-\lambda M_{\mathrm{T}}})} \sum_{k=1}^{N}\left(1-e^{-\lambda M'_{g_{k,n}}}\right)\right). \tag{B.21}$$

We next derive $\mathbb{E}[\zeta_j^2]$. Based on the independent and identical distribution of B_{j,h_1} and B_{j,h_2} when $h_1 \neq h_2$, we calculate $\mathbb{E}\left[\zeta_j^2\right]$ from (B.10) as

$$\begin{aligned}\mathbb{E}\left[\zeta_j^2\right] =& \frac{1}{1-\varepsilon}\left(M_{\mathrm{T}}^2 + 2M_{\mathrm{T}}\mathbb{E}[V_{j,h}] + \mathbb{E}\left[V_{j,h}^2\right]\right) + \frac{2\varepsilon}{(1-\varepsilon)^2}(M_{\mathrm{T}}+\xi)^2 \\ =& \frac{1+\varepsilon}{(1-\varepsilon)^2}(M_{\mathrm{T}}+\xi)^2 + \frac{2\xi-\lambda\xi^2}{\lambda(1-\varepsilon)},\end{aligned} \tag{B.22}$$

where $\mathbb{E}[V_{j,h}^2] = \frac{2}{\lambda^2}\sum_{k=1}^{N}\Pr(Z_{j,k,h})$ and $\mathbb{E}[V_{j,h}]$ is given in (B.20). By substituting (B.18), (B.21), and (B.22) into (5.6), we obtain the final result given in (5.9), which completes this proof.

B.3 Proof of Theorem 5.3

Based on (5.6), we need to calculate $\mathbb{E}[T_j]$, $\mathbb{E}[\zeta_j]$, and $\mathbb{E}[\zeta_j^2]$ to obtain $\Delta_{\mathrm{UNIPS},n}$. We first derive $\mathbb{E}[T_j]$. In the UNIPS strategy, the waiting time of P_j is 0, i.e., $W_j = 0$, since preemption in serving is considered. Thus, we calculate $\mathbb{E}[T_j]$ as

$$\mathbb{E}[T_j] = \mathbb{E}[S_j]. \tag{B.23}$$

We note that in this strategy, the probability of $A_{j,k}$ and $G_{j,k}$ are given by

$$\Pr\left(A_{j,k}\right) = \begin{cases} \prod_{\kappa=k}^{N} p_k(1-p_\kappa), & \text{if } k \neq 0, \\ \prod_{\kappa=1}^{N}(1-p_\kappa), & \text{otherwise}, \end{cases} \tag{B.24}$$

and

$$\Pr\left(G_{j,k}\right) = \frac{1}{1-e^{-\lambda M_{\mathrm{T}}}}\Pr\left(A_{j,k}\right), \tag{B.25}$$

respectively. Based on (B.25), we calculate $\mathbb{E}[T_j]$ as

$$\mathbb{E}[T_j] = \mathbb{E}[S_j] = \sum_{k=1}^{N}\Pr\left(G_{j,k}\right)\sum_{\kappa=k}^{N} M'_{g_{\kappa,n}} = \sum_{k=1}^{N} M'_{g_{k,n}} \frac{e^{-\lambda\sum_{\kappa=k+1}^{N} M'_{g_{\kappa,n}}} - e^{-\lambda M_{\mathrm{T}}}}{1-e^{-\lambda M_{\mathrm{T}}}}. \tag{B.26}$$

We then derive $\mathbb{E}[\zeta_j]$ based on (B.3). We note that

$$B_{j,h} = V_{j,h} + \sum_{k=1}^{N} S_{j,k,h}, \tag{B.27}$$

where $S_{j,k,h}$ is the serving time of the $U_{g_{k,n}}$'s packet during $B_{j,h}$. We then note that the probability of $Z_{j,k,h}$ is given by

$$\Pr\left(Z_{j,k,h}\right) = \frac{(1-e^{-\lambda M'_{g_{k+1,n}}})e^{-\lambda M_{\mathrm{T}}}}{1-e^{-\lambda M_{\mathrm{T}}}}, \tag{B.28}$$

for the UNIPS strategy. Based on (B.28), we calculate $\mathbb{E}\left[V_{j,h}\right]$ as

$$\mathbb{E}\left[V_{j,h}\right] = \frac{1}{\lambda}\sum_{k=1}^{N}\Pr\left(Z_{j,k,h}\right) = \frac{e^{-\lambda M_{\mathrm{T}}}}{\lambda\left(1-e^{-\lambda M_{\mathrm{T}}}\right)}\sum_{k=1}^{N}\frac{p_k}{1-p_k}. \tag{B.29}$$

For $S_{j,k,h}$, it is calculated as

$$S_{j,k,h} = \sum_{\theta=0}^{\Theta_k} S_{j,k,h,\theta} + M'_{g_{k,n}}, \tag{B.30}$$

where Θ_k is the times of preemption of status update for $U_{g_{k,n}}$ and $S_{j,k,h,\theta}$ is the serving time of the θth preempted status update. We note that Θ_k follows a geometric distribution, i.e., $\Pr\left(\Theta_k = \Theta\right) = (1-p_k)p_k^{\Theta}$, and the average serving time of the θth preempted status update is calculated as

$$\mathbb{E}\left[S_{j,k,h,\theta}\right] = \frac{1}{p_k}\int_0^{M'_{g_{k,n}}} \lambda t e^{-\lambda t}\mathrm{d}t = \frac{1}{\lambda} - \varphi, \tag{B.31}$$

where $\varphi = \frac{(1-p_k)M'_{g_{k,n}}}{p_k}$. Combining (B.30) with (B.31), we calculate $\mathbb{E}\left[S_{j,k,h}\right]$ by

$$\mathbb{E}\left[S_{j,k,h}\right] = \mathbb{E}\left[\Theta_k\right]\mathbb{E}[S_{j,k,h,\theta}] + M'_{g_{k,n}} = \frac{p_k}{1-p_k}\left(\frac{1}{\lambda} - \varphi\right) + M'_{g_{k,n}} = \frac{p_k}{\lambda(1-p_k)}. \tag{B.32}$$

By substituting (B.29) and (B.32) into (B.27), we obtain $\mathbb{E}[B_{j,h}]$ as

$$\mathbb{E}[B_{j,h}] = \frac{1}{\lambda(1-e^{-\lambda M_{\mathrm{T}}})} \sum_{k=1}^{N} \frac{p_k}{1-p_k}. \tag{B.33}$$

Then, by substituting (B.33) into (B.3), we obtain $\mathbb{E}[\zeta_j]$ as

$$\mathbb{E}[\zeta_j] = \frac{1}{\lambda(1-\varepsilon_n)(1-e^{-\lambda M_{\mathrm{T}}})} \sum_{k=1}^{N} \frac{p_k}{1-p_k}. \tag{B.34}$$

We next calculate $\mathbb{E}\left[\zeta_j^2\right]$ based on (B.10). According to (B.27), we calculate $\mathbb{E}\left[B_{j,h}^2\right]$ in (B.10) as

$$\mathbb{E}\left[B_{j,h}^2\right] = \mathbb{E}\left[V_{j,h}^2\right] + 2\mathbb{E}\left[V_{j,h}\left(\sum_{k=1}^{N} S_{j,k,h}\right)\right] + \mathbb{E}\left[\left(\sum_{k=1}^{N} S_{j,k,h}\right)^2\right]. \tag{B.35}$$

We note that $V_{j,h}$ is not independent from $S_{j,k,h}$ in (B.35). If the BS is in the idle state before transmitting the packet of $U_{g_{k,n}}$, the packet of $U_{g_{\kappa,n}}$, $\kappa \in \{1,2,\cdots,k\}$, is not preempted during $B_{j,h}$. Hence, we obtain

$$\mathbb{E}\left[V_{j,h}\left(\sum_{k=1}^{N} S_{j,k,h}\right)\right] = \frac{e^{-\lambda M'_{\mathrm{T}}}}{\lambda(1-e^{-\lambda M'_{\mathrm{T}}})} \sum_{k=1}^{N} \frac{p_k}{1-p_k}\left(\sum_{\kappa=k}^{N} \frac{p_\kappa}{\lambda(1-p_\kappa)}\right). \tag{B.36}$$

We then calculate $\mathbb{E}\left[V_{j,h}^2\right]$ and $\mathbb{E}\left[\left(\sum_{k=1}^{N} S_{j,k,h}\right)^2\right]$ in (B.35) as

$$\mathbb{E}\left[V_{j,h}^2\right] = \frac{2e^{-\lambda M_{\mathrm{T}}}}{\lambda^2(1-e^{-\lambda M_{\mathrm{T}}})} \sum_{k=1}^{N} \frac{p_k}{1-p_k}, \tag{B.37}$$

and

$$\mathbb{E}\left[\left(\sum_{k=1}^{N} S_{j,k,h}\right)^2\right] = \sum_{k=1}^{N}\frac{p_k}{1-p_k}\left(\frac{2}{\lambda^2}-\frac{2\varphi}{\lambda}-\varphi M'_{g_{k,n}}\right)$$
$$+\sum_{k=1}^{N}\sum_{\kappa\neq k}\frac{p_k}{1-p_k}\left(\frac{1}{\lambda}-\varphi\right)\frac{p_\kappa}{1-p_\kappa}\left(\frac{1}{\lambda}-\frac{M'_{g_{\kappa,n}}(1-p_\kappa)}{p_\kappa}\right)$$
$$+\sum_{k=1}^{N}\frac{2p_k^2}{(1-p_k)^2}\left(\frac{1}{\lambda}-\varphi\right)^2+2M_{\mathrm{T}}\sum_{k=1}^{N}\left(\frac{p_k}{1-p_k}\left(\frac{1}{\lambda}-\varphi\right)\right)+M_{\mathrm{T}}^2, \tag{B.38}$$

respectively. By substituting (B.36), (B.37), and (B.38) into (B.35), and combining (B.35) with (B.10), we obtain $\mathbb{E}\left[\zeta_j^2\right]$. By substituting (B.26), (B.34), and $\mathbb{E}\left[\zeta_j^2\right]$ into (5.6), we obtain the final result in (5.12), which completes the proof.

B.4 Proof of Theorem 5.4

In the remote control system, we find that the block error rate ε_n is negligible, as a low coding rate is assumed. With the information ratio, α, and the same coding rate, R, we obtain the transmitted packet length, M, in the BRNP strategy as

$$M=\frac{L}{R}=\frac{\alpha N L_h}{R}=\alpha\rho M_{\mathrm{T}}. \tag{B.39}$$

Hence, the average AoI in the BRNP strategy is calculated as

$$\begin{aligned}\Delta_{\mathrm{BRNP}} &= \frac{1}{2}\left(M+\frac{1}{\lambda}e^{-\lambda M}\right)+\frac{2e^{-\lambda M}-e^{-2\lambda M}}{2\lambda^2(M+\frac{1}{\lambda}e^{-\lambda M})}+\left(\frac{1}{\lambda}+M\right)\left(1-e^{-\lambda M}\right)\\ &=\frac{1}{2}\left(\alpha\rho M_{\mathrm{T}}+\frac{1}{\lambda}e^{-\lambda\alpha\rho M_{\mathrm{T}}}\right)+\frac{2e^{-\lambda\alpha\rho M_{\mathrm{T}}}-e^{-2\lambda\alpha\rho M_{\mathrm{T}}}}{2\lambda^2(\alpha\rho M_{\mathrm{T}}+\frac{1}{\lambda}e^{-\lambda\alpha\rho M_{\mathrm{T}}})}+\left(\frac{1}{\lambda}+\alpha\rho M_{\mathrm{T}}\right)\left(1-e^{-\lambda\alpha\rho M_{\mathrm{T}}}\right).\end{aligned} \tag{B.40}$$

Since all UEs require the same length of the status update, L_h, and have the same coding rate R, the system average AoI in the UNINP strategy is calculated as

$$\Delta_{\mathrm{UNINP}}=\frac{1}{2}\left(M_{\mathrm{T}}+\frac{1}{\lambda}e^{-\lambda M_{\mathrm{T}}}\right)+\frac{2e^{-\lambda M_{\mathrm{T}}}-e^{-2\lambda M_{\mathrm{T}}}}{2\lambda^2(M_{\mathrm{T}}+\frac{1}{\lambda}e^{-\lambda M_{\mathrm{T}}})}-\frac{N-1}{2N}M_{\mathrm{T}}+\left(\frac{1}{\lambda}+M_{\mathrm{T}}\right)(1-e^{-\lambda M_{\mathrm{T}}}). \tag{B.41}$$

By calculating the average AoI difference between the BRNP strategy and the UNINP strategy, we obtain the final result in (5.15).

B.5 Proof of Theorem 5.5

In the dynamic system, we denote N_Λ as the number of UEs, which follows a Poisson distribution with the PMF given by

$$\Pr(N_\Lambda = N) = \frac{\Lambda^N e^{-\Lambda}}{N!}. \tag{B.42}$$

We first analyze the expected average AoI in the BRNP strategy. We assume that the coding rate of transmission is close to the channel capacity and the block error rate is negligible. Then, the average AoI of the system with N UEs, where $N \neq 0$, is approximated as

$$\Delta_{\mathrm{BRNPZ}}^{(N)} \approx \frac{3(L_{\mathrm{co}} + NL_{\mathrm{id}})}{2C_{D_2}}. \tag{B.43}$$

By averaging the average AoI w.r.t. the number of UEs, the expected average AoI is calculated as

$$\mathbb{E}[\Delta_{\mathrm{BRNPZ}}] = \sum_{N=1}^{\infty} \Pr(N_\Lambda = N)\Delta_{\mathrm{BRNPZ}}^{(N)} \approx \sum_{N=1}^{\infty} \frac{3\Lambda^N e^{-\Lambda}(L_{\mathrm{co}} + NL_{\mathrm{id}})}{2N!C_{D_2}}, \tag{B.44}$$

which results in (5.20).

We then analyze the expected average AoI in the UNINP strategy. Here, we approximate the average AoI by considering the coding rate to be close to the channel capacity of each UE and neglecting the block error rate. Hence, we approximate the average AoI of the system with N UEs, where $N \neq 0$, as

$$\Delta_{\mathrm{UNINPZ}}^{(N)} \approx \frac{3}{2}\sum_{n=1}^{N}\left(\mathbb{E}\left[\frac{L_{\mathrm{co}} + L_{\mathrm{id}}}{C_n}\right] + M_{\mathrm{L}}\right) - \frac{1}{N}\sum_{n=1}^{N}\sum_{k=n+1}^{N}\left(\mathbb{E}\left[\frac{L_{\mathrm{co}} + L_{\mathrm{id}}}{C_n}\right] + M_{\mathrm{L}}\right), \tag{B.45}$$

where C_n is the channel capacity of U_n. Due to the fact that the location of UEs follows the same Poisson process, we simplify (B.45) as

$$\Delta_{\mathrm{UNINPZ}}^{(N)} \approx \frac{2N+1}{2}\left((L_{\mathrm{co}} + L_{\mathrm{id}})\,\mathbb{E}\left[\frac{1}{C_n}\right] + M_{\mathrm{L}}\right). \tag{B.46}$$

To obtain $\Delta_{\mathrm{UNINPZ}}^{(N)}$, we need to calculate $\mathbb{E}[1/C_n]$. Since the distance from each UE to the BS follows a 2D Poisson process, the PDF of the distance between the BS and each UE is given by

$$f(d) = \frac{2d}{D_2^2 - D_1^2}. \tag{B.47}$$

With high received SNR, the channel capacity is approximated as

$$C_n = \frac{1}{2}\log_2\left(1+\gamma_n\right) \approx \frac{1}{2}\log_2\left(\gamma_n\right). \tag{B.48}$$

Based on (B.47) and (B.48), we calculate $\mathbb{E}\left[1/C_n\right]$ as

$$\mathbb{E}\left[\frac{1}{C_n}\right] \approx \int_{D_1}^{D_2} \frac{1}{C_0 - \frac{\eta}{2}\log_2 d} f(d)\mathrm{d}d = \frac{2^{2+\frac{4C_0}{\eta}}\ln 2}{\left(D_2^2 - D_1^2\right)\eta}\left(\mathrm{Ei}\left(x_1\right) - \mathrm{Ei}\left(x_2\right)\right) = \frac{1}{C_\Lambda}. \tag{B.49}$$

By substituting (B.49) into (B.46), we obtain the average AoI of the system with N UEs in the UNINP strategy as

$$\Delta_{\mathrm{UNINPZ}}^{(N)} \approx \frac{2N+1}{2}\left(\frac{L_{\mathrm{co}}+L_{\mathrm{id}}}{C_\Lambda} + M_{\mathrm{L}}\right). \tag{B.50}$$

By averaging the average AoI w.r.t. the number of UEs, the expected average AoI is calculated as

$$\mathbb{E}[\Delta_{\mathrm{UNINPZ}}] = \sum_{N=1}^{\infty} \Pr(N_\Lambda = N)\Delta_{\mathrm{UNINPZ}}^{(N)} \approx \sum_{N=1}^{\infty} \frac{\Lambda^N e^{-\Lambda}\left(2N+1\right)}{2N!}\left(\frac{L_{\mathrm{co}}+L_{\mathrm{id}}}{C_n} + M_{\mathrm{L}}\right), \tag{B.51}$$

which results in (5.21).

Bibliography

[1] Ericsson, “Ericsson mobility report,” pp. 1 –40, Jun. 2023. [Online]. Available: https://www.ericsson.com/49dd9d/assets/local/reports-papers/mobility-report/documents/2023/ericsson-mobility-report-june-2023.pdf (cited on page 1)

[2] ITU-R, “Framework and overall objectives of the future development of IMT for 2030 and beyond,” Jun. 2023. (cited on page 1)

[3] C. Yan, Z. Peiying, and T. Wen, “ITU-R WP5D completed the recommendation framework for IMT-2030 (global 6G vision),” Jun. 2023. [Online]. Available: https://www.huawei.com/en/huaweitech/future-technologies/itu-r-wp5d-completed-recommendation-framework-imt-2030#text1 (cited on page 1)

[4] A. Varghese and D. Tandur, “Wireless requirements and challenges in industry 4.0,” in *Proc. IEEE Int. Conf. Contemp. Comput. Informat.*, Mysore, India, Nov. 2014, pp. 634–638. (cited on page 2)

[5] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, “5G-enabled tactile Internet,” *IEEE J. Select. Areas Commun.*, vol. 34, no. 3, pp. 460–473, Mar. 2016. (cited on page 2)

[6] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, “Fundamental tradeoffs among reliability, latency and throughput in cellular networks,” in *Proc. IEEE Global Commun. Conf. Workshop*, Austin, TX, USA, Dec. 2014, pp. 1391–1396. (cited on page 3)

[7] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, “Minimizing age of information in vehicular networks,” in *Proc. IEEE Conf. Sensor Ad Hoc Commun. Netw.*, Salt Lake City, UT, USA, Jun. 2011, pp. 350–358. (cited on pages 3 and 7)

[8] L. Ying and S. Shakkottai, “On throughput optimality with delayed network-state information,” *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5116–5132, Aug. 2011. (cited on page 5)

[9] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, “Age and value of information:

Non-linear age case," in *Proc. IEEE Int. Sympos. Inf. Theory*, Aachen, Germany, Jun. 2017, pp. 326–330. (cited on page 6)

[10] ——, "The cost of delay in status updates and their value: Non-linear ageing," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4905 – 4918, Apr. 2020. (cited on page 6)

[11] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Minimizing the age of information through queues," *IEEE Trans. Inf. Theory*, vol. 65, no. 8, pp. 5215–5232, Aug. 2019. (cited on page 6)

[12] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE Int. Conf. Comput. Commun.*, Orlando, FL, USA, Mar. 2012, pp. 2731–2735. (cited on pages 7, 8, and 47)

[13] J. P. Champati, H. Al-Zubaidy, and J. Gross, "Statistical guarantee optimization for age of information for the D/G/1 queue," in *Proc. IEEE Int. Conf. Comput. Commun. Workshop*, Honolulu, HI, USA, Apr. 2018, pp. 130–135. (cited on page 7)

[14] A. Soysal and S. Ulukus, "Age of information in G/G/1/1 systems: Age expressions, bounds, special cases, and optimization," *IEEE Trans. Inf. Theory*, vol. 67, no. 11, pp. 7477–7489, Nov. 2021. (cited on page 7)

[15] C. Kam, S. Kompella, G. D. Nguyen, and A. Ephremides, "Effect of message transmission path diversity on status age," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1360–1374, Mar. 2016. (cited on page 7)

[16] R. D. Yates and S. Kaul, "Real-time status updating: Multiple sources," in *Proc. IEEE Int. Sympos. Inf. Theory*, Cambridge, MA, USA, Jul. 2012, pp. 2666–2670. (cited on page 8)

[17] L. Huang and E. Modiano, "Optimizing age-of-information in a multi-class queueing system," in *Proc. IEEE Int. Sympos. Inf. Theory*, Hong Kong, China, Oct. 2015, pp. 1681–1685. (cited on page 8)

[18] R. D. Yates and S. K. Kaul, "The age of information: Real-time status updating by multiple sources," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1807–1827, Mar. 2019. (cited on pages 8, 45, 65, and 66)

[19] S. K. Kaul, R. D. Yates, and M. Gruteser, "Status updates through queues," in *Proc. Conf. Inf. Sciences Systems*, Baltimore, MD, USA, Mar. 2012, pp. 1–6. (cited on pages 8 and 88)

[20] M. Costa, M. Codreanu, and A. Ephremides, "Age of information with packet management," in *Proc. IEEE Int. Sympos. Inf. Theory*, Honolulu, HI, USA, Jun. 2014, pp. 1583–1587. (cited on page 8)

[21] R. D. Yates, "The age of information in networks: Moments, distributions, and sampling," *IEEE Trans. Inf. Theory*, vol. 66, no. 9, pp. 5712–5728, Sep. 2020. (cited on page 8)

[22] R. Talak, S. Karaman, and E. Modiano, "Minimizing age-of-information in multi-hop wireless networks," in *Proc. Allerton Conf. Commun. Control Comput.*, Monticello, IL, USA, Oct. 2017, pp. 486–493. (cited on page 8)

[23] R. D. Yates, "Age of information in a network of preemptive servers," in *Proc. IEEE Int. Conf. Comput. Commun. Workshop*, Honolulu, HI, USA, Apr. 2018, pp. 118–123. (cited on page 8)

[24] Z. Jiang, B. Krishnamachari, X. Zheng, S. Zhou, and Z. Niu, "Timely status update in wireless uplinks: Analytical solutions with asymptotic optimality," *IEEE Intern. of Things J.*, vol. 6, no. 2, pp. 3885–3898, Apr. 2019. (cited on page 8)

[25] K. Zheng, S. Ou, J. Alonso-Zarate, M. Dohler, F. Liu, and H. Zhu, "Challenges of massive access in highly dense LTE-advanced networks with machine-to-machine communications," *IEEE Wireless Commun.*, vol. 21, no. 3, pp. 12–18, Jun. 2014. (cited on page 8)

[26] A. Gong, T. Zhang, H. Chen, and Y. Zhang, "Age-of-information-based scheduling in multiuser uplinks with stochastic arrivals: A POMDP approach," in *Proc. IEEE Global Commun. Conf.*, Taipei, Taiwan, China, Dec. 2020. (cited on page 8)

[27] Y. Shao, Q. Cao, and S. C. Liew, "Partially observable minimum-age scheduling: The greedy policy," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 404–418, Jan. 2022. (cited on page 8)

[28] I. Kadota, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Minimizing the age of information in broadcast wireless networks," in *Proc. Allerton Conf. Commun. Control Comput.*, Monticello, IL, USA, Sep. 2016, pp. 844–851. (cited on pages 8 and 9)

[29] Y. Hsu, "Age of information: Whittle index for scheduling stochastic arrivals," in *Proc. IEEE Int. Sympos. Inf. Theory*, Vail, CO, USA, Jun. 2018, pp. 2634–2638. (cited on pages 9 and 107)

[30] J. Sun, Z. Jiang, S. Zhou, and Z. Niu, "Optimizing information freshness in broadcast network with unreliable links and random arrivals: An approximate index policy," in

Proc. IEEE Int. Conf. Comput. Commun. Workshop, Paris, France, Apr. 2019, pp. 115–120. (cited on pages 9, 33, and 39)

[31] J. Sun, Z. Jiang, B. Krishnamachari, S. Zhou, and Z. Niu, "Closed-form Whittle's index-enabled random access for timely status update," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1538–1551, Dec. 2020. (cited on pages 9 and 107)

[32] A. Elgabli, H. Khan, M. Krouka, and M. Bennis, "Reinforcement learning based scheduling algorithm for optimizing age of information in ultra reliable low latency networks," in *Proc. IEEE Symp. Comput. Commun.*, Barcelona, Spain, Jun. 2019, pp. 1–6. (cited on page 9)

[33] H. B. Beytur and E. Uysal, "Age minimization of multiple flows using reinforcement learning," in *Proc. Int. Conf. Comput., Netw. Commun.*, Honolulu, HI, USA, Feb. 2019, pp. 339–343. (cited on page 9)

[34] M. Bastopcu and S. Ulukus, "Who should google scholar update more often?" in *Proc. IEEE Int. Conf. Comput. Commun. Workshop*, Toronto, ON, Canada, Jul. 2020, pp. 696–701. (cited on pages 9 and 107)

[35] B. Yin, S. Zhang, and Y. Cheng, "Application-oriented scheduling for optimizing the age of correlated information: A deep-reinforcement-learning-based approach," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8748–8759, 2020. (cited on page 9)

[36] E. T. Ceran, D. Gündüz, and A. György, "A reinforcement learning approach to age of information in multi-user networks with HARQ," *IEEE J. Select. Areas Commun.*, vol. 39, no. 5, pp. 1412–1426, May 2021. (cited on pages 9 and 107)

[37] M. Klügel, M. H. Mamduhi, S. Hirche, and W. Kellerer, "AoI-penalty minimization for networked control systems with packet loss," in *Proc. IEEE Int. Conf. Comput. Commun.*, Paris, France, May 2019, pp. 189–196. (cited on page 9)

[38] V. Tripathi and E. Modiano, "A Whittle index approach to minimizing functions of age of information," in *Proc. Allerton Conf. Commun. Control, Comput.*, Monticello, IL, USA, Sep. 2019, pp. 1160–1167. (cited on pages 9, 39, and 107)

[39] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing-A key technology towards 5G," *ETSI white paper*, vol. 11, no. 11, pp. 1–16, Sep. 2015. (cited on page 9)

[40] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, Aug. 2017. (cited on page 9)

[41] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, Mar. 2017. (cited on page 9)

[42] Y. Zhang, H. Liu, L. Jiao, and X. Fu, "To offload or not to offload: An efficient code partition algorithm for mobile cloud computing," in *Proc. IEEE Intern. Conf. Cloud Netw.*, Paris, France, Nov. 2012, pp. 80–86. (cited on page 9)

[43] X. Chen, "Decentralized computation offloading game for mobile cloud computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 4, pp. 974–983, Apr. 2015. (cited on page 10)

[44] J. Liu, Y. Mao, J. Zhang, and K. B. Letaief, "Delay-optimal computation task scheduling for mobile-edge computing systems," in *Proc. IEEE Int. Sympos. Inf. Theory*, Barcelona, Spain, Jul. 2016, pp. 1451–1455. (cited on page 10)

[45] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Select. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, Sep. 2016. (cited on page 10)

[46] L. Zhao, J. Wang, J. Liu, and N. Kato, "Optimal edge resource allocation in IoT-based smart cities," *IEEE Netw.*, vol. 33, no. 2, pp. 30–35, Mar. 2019. (cited on page 10)

[47] J. Ren, G. Yu, Y. He, and G. Y. Li, "Collaborative cloud and edge computing for latency minimization," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 5031–5044, May 2019. (cited on page 10)

[48] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015. (cited on page 10)

[49] C. You, K. Huang, H. Chae, and B. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1397–1411, Mar. 2017. (cited on page 10)

[50] T. G. Rodrigues, K. Suto, H. Nishiyama, N. Kato, and K. Temma, "Cloudlets activation scheme for scalable mobile edge computing with transmission power control and virtual machine migration," *IEEE Trans. Comput.*, vol. 67, no. 9, pp. 1287–1300, Sep. 2018. (cited on page 10)

[51] B. Zhou and W. Saad, "Optimal sampling and updating for minimizing age of information in the internet of things," in *Proc. IEEE Global Commun. Conf.*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6. (cited on page 10)

[52] R. Li, Q. Ma, J. Gong, Z. Zhou, and X. Chen, “Age of processing: Age-driven status sampling and processing offloading for edge-computing-enabled real-time IoT applications,” *IEEE Internet Things J.*, vol. 8, no. 19, pp. 14 471–14 484, Oct. 2021. (cited on page 10)

[53] Q. Kuang, J. Gong, X. Chen, and X. Ma, “Analysis on computation-intensive status update in mobile edge computing,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4353–4366, Apr. 2020. (cited on pages 10 and 45)

[54] P. Zou, O. Ozel, and S. Subramaniam, “Optimizing information freshness through computation–transmission tradeoff and queue management in edge computing,” *IEEE/ACM Trans. on Netw.*, vol. 29, no. 2, pp. 949–963, Oct. 2021. (cited on page 10)

[55] A. Arafa, R. D. Yates, and H. V. Poor, “Timely cloud computing: Preemption and waiting,” in *Proc. Allerton Conf. Commun. Control Comput.*, Monticello, IL, USA, Sep. 2019, pp. 528–535. (cited on page 10)

[56] B. Buyukates and S. Ulukus, “Timely distributed computation with stragglers,” *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5273–5282, Sep. 2020. (cited on page 10)

[57] C. Xu, H. Yang, X. Wang, and T. Q. Quek, “Optimizing information freshness in computing-enabled IoT networks,” *IEEE Internet Things J.*, vol. 7, no. 2, pp. 971–985, Feb. 2020. (cited on page 10)

[58] B. Zhou and W. Saad, “Minimum age of information in the internet of things with non-uniform status packet sizes,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1933–1947, Mar. 2020. (cited on page 10)

[59] H. Wu, H. Tian, S. Fan, and J. Ren, “Data age aware scheduling for wireless powered mobile-edge computing in industrial internet of things,” *IEEE Trans. Ind. Informat*, vol. 17, no. 1, pp. 398–408, Apr. 2020. (cited on page 10)

[60] G. Durisi, T. Koch, and P. Popovski, “Toward massive, ultrareliable, and low-latency wireless communication with short packets,” *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016. (cited on page 11)

[61] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, “Short-packet communications over multiple-antenna Rayleigh-fading channels,” *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 618–629, Feb. 2016. (cited on page 11)

[62] Y. Polyanskiy, H. V. Poor, and s. Verdu, “Channel coding rate in the finite blocklength regime,” *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, Apr. 2010. (cited on pages 11, 63, 73, and 85)

[63] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal-Biyikoglu, "Delay and peak-age violation probability in short-packet transmissions," in *Proc. IEEE Int. Sympos. Inf. Theory*, Vail, CO, USA, Jun. 2018, pp. 2471–2475. (cited on pages 11, 61, and 85)

[64] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal, "Reliable transmission of short packets through queues and noisy channels under latency and peak-age violation guarantees," *IEEE J. Select. Areas Commun.*, vol. 37, no. 4, pp. 721–734, Apr. 2019. (cited on pages 11, 61, and 85)

[65] J. Ostman, R. Devassy, G. Durisi, and E. Uysal, "Peak-age violation guarantees for the transmission of short packets over fading channels," in *Proc. IEEE Int. Conf. Comput. Commun.*, Paris, France, May 2019, pp. 109–114. (cited on pages 11 and 61)

[66] C. M. W. Basnayaka, D. N. K. Jayakody, T. D. P. Perera, and M. V. Ribeiro, "Age of information in an URLLC-enabled decode-and-forward wireless communication system," in *Proc. IEEE Veh. Techn. Conf.*, Helsinki, Finland, Apr. 2021, pp. 1–6. (cited on pages 11 and 61)

[67] R. Wang, Y. Gu, H. Chen, Y. Li, and B. Vucetic, "On the age of information of short-packet communications with packet management," in *Proc. IEEE Global Commun. Conf.*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6. (cited on pages 11, 61, 63, 73, and 85)

[68] H. Sung, M. Kim, S. Lee, and J. Lee, "Age of information analysis for finite blocklength regime in downlink cellular networks," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 683–687, Apr. 2022. (cited on pages 11 and 61)

[69] S. Roth, A. Arafa, A. Sezgin, and H. V. Poor, "Short blocklength process monitoring and scheduling: Resolution and data freshness," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 4669–4681, Dec 2022. (cited on pages 11 and 61)

[70] B. Yu, Y. Cai, D. Wu, and Z. Xiang, "Average age of information in short packet based machine type communication," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10 306–10 319, Jun. 2020. (cited on pages 11 and 61)

[71] A. Arafa and R. D. Wesel, "Timely transmissions using optimized variable length coding," in *Proc. IEEE Int. Conf. Comput. Commun. Workshop*, Vancouver, BC, Canada, May 2021, pp. 1–6. (cited on pages 11 and 61)

[72] A. Arafa, K. Banawan, K. G. Seddik, and H. V. Poor, "On timely channel coding with hybrid ARQ," in *Proc. IEEE Global Commun. Conf.*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6. (cited on pages 11 and 61)

[73] B. Li, Z. Fei, and Y. Zhang, "UAV communications for 5G and beyond: Recent advances and future trends," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2241–2263, Apr. 2019. (cited on page 11)

[74] P. Zhan, K. Yu, and A. L. Swindlehurst, "Wireless relay communications with unmanned aerial vehicles: Performance and optimization," *IEEE Trans Aerosp Electron Syst.*, vol. 47, no. 3, pp. 2068–2085, Jul. 2011. (cited on page 12)

[75] C. She, C. Liu, T. Q. S. Quek, C. Yang, and Y. Li, "UAV-assisted uplink transmission for ultra-reliable and low-latency communications," in *Proc. IEEE Int. Commun. Conf. Workshop*, Kansas City, MO, May 2018, pp. 1–6. (cited on page 12)

[76] S. Zhang, H. Zhang, Z. Han, H. V. Poor, and L. Song, "Age of information in a cellular internet of UAVs: Sensing and communication trade-off design," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6578–6592, Oct. 2020. (cited on page 12)

[77] C. M. W. Basnayaka, D. N. K. Jayakody, and Z. Chang, "Age-of-information-based urllc-enabled UAV wireless communications system," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 10 212–10 223, Jun. 2022. (cited on page 12)

[78] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017. (cited on page 12)

[79] A. C. Pogaku, D.-T. Do, B. M. Lee, and N. D. Nguyen, "UAV-assisted RIS for future wireless communications: A survey on optimization and performance analysis," *IEEE Access*, vol. 10, pp. 16 320–16 336, Feb. 2022. (cited on page 12)

[80] M. Samir, M. Elhattab, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Optimizing age of information through aerial reconfigurable intelligent surfaces: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3978–3983, Apr. 2021. (cited on page 12)

[81] B. Yu, X. Guan, and Y. Cai, "Joint blocklength and power optimization for half duplex unmanned aerial vehicle relay system with short packet communications," in *Proc. Int. Conf. Wireless Commun. Signal Process.*, Nanjing, China, Oct. 2020, pp. 981–986. (cited on page 12)

[82] J. Liu, X. Wang, B. Bai, and H. Dai, "Age-optimal trajectory planning for UAV-assisted data collection," in *Proc. IEEE Int. Conf. Comput. Commun.*, Honolulu, HI, Apr. 2018, pp. 553–558. (cited on page 12)

[83] Y. Luo, J. Xu, J. Chen, and J. Huang, "UAV trajectory planning with network age of information minimization," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Austin, TX, May 2022, pp. 1862–1867. (cited on page 12)

[84] P. Tong, J. Liu, X. Wang, B. Bai, and H. Dai, "UAV-enabled age-optimal data collection in wireless sensor networks," in *Proc. IEEE Int. Commun. Conf. Workshop*, Shanghai, China, May 2019, pp. 1–6. (cited on page 12)

[85] G. Ahani, D. Yuan, and Y. Zhao, "Age-optimal UAV scheduling for data collection with battery recharging," *IEEE Commun. Lett.*, vol. 25, no. 4, pp. 1254–1258, Apr. 2021. (cited on page 12)

[86] C. Mao, J. Liu, and L. Xie, "Multi-UAV aided data collection for age minimization in wireless sensor networks," in *Proc. Int. Conf. Wireless Commun. Signal Process.*, Nanjing, China, Oct. 2020, pp. 80–85. (cited on page 12)

[87] C. Zhao, J. Liu, M. Sheng, W. Teng, Y. Zheng, and J. Li, "Multi-UAV trajectory planning for energy-efficient content coverage: A decentralized learning-based approach," *IEEE J. Select. Areas Commun.*, vol. 39, no. 10, pp. 3193–3207, Oct. 2021. (cited on page 12)

[88] M. A. Abd-Elmagid and H. S. Dhillon, "Age of information in multi-source updating systems powered by energy harvesting," *IEEE J. Select. Areas Commun.*, vol. 3, no. 1, pp. 98–112, Mar. 2022. (cited on page 12)

[89] Q. Dang, Q. Cui, Z. Gong, X. Zhang, X. Huang, and X. Tao, "AoI oriented UAV trajectory planning in wireless powered IoT networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Austin, TX, May 2022, pp. 884–889. (cited on page 12)

[90] H. Hu, K. Xiong, G. Qu, Q. Ni, P. Fan, and K. B. Letaief, "AoI-minimal trajectory planning and data collection in UAV-assisted wireless powered IoT networks," *IEEE Internet of Things Journal*, vol. 8, no. 2, pp. 1211–1223, Jan. 2021. (cited on page 12)

[91] J. Hu, H. Zhang, L. Song, R. Schober, and H. V. Poor, "Cooperative internet of UAVs: Distributed trajectory design by multi-agent deep reinforcement learning," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 6807–6821, Nov. 2020. (cited on page 12)

[92] W. Li, L. Wang, and A. Fei, "Minimizing packet expiration loss with path planning in UAV-assisted data sensing," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1520–1523, Dec. 2019. (cited on page 12)

[93] F. Wu, H. Zhang, J. Wu, L. Song, Z. Han, and H. V. Poor, "AoI minimization for UAV-to-device underlay communication by multi-agent deep reinforcement learning,"

in *Proc. IEEE Global Commun. Conf.*, Taipei, Taiwan, Dec. 2020, pp. 1–6. (cited on page 12)

[94] Z. Han, J. Liang, Y. Gu, and H. Chen, "Software-defined radio implementation of age-of-information-oriented random access," in *Proc. Conf. IEEE Ind. Electron. Soc.*, Singapore, Oct. 2020, pp. 4374–4379. (cited on page 13)

[95] O. Ayan, H. Y. Özkan, and W. Kellerer, "An experimental framework for age of information and networked control via software-defined radios," in *Proc. IEEE Int. Commun. Conf.*, Montreal, QC, Canada, Jun. 2021, pp. 1–6. (cited on page 13)

[96] I. Kadota, M. S. Rahman, and E. Modiano, "Wifresh: Age-of-information from theory to implementation," in *Proc. IEEE Int. Conf. Comput. Commun. Netw.*, Athens, Greece, Jul. 2021, pp. 1–11. (cited on page 13)

[97] "IEEE standard for information technology–telecommunications and information exchange between systems - local and metropolitan area networks–specific requirements - part 11: Wireless lan medium access control (MAC) and physical layer (PHY) specifications," *IEEE Std 802.11-2020 (Revision of IEEE Std 802.11-2016)*, pp. 1–4379, Feb. 2021. (cited on page 13)

[98] Z. Tang, Z. Sun, N. Yang, and X. Zhou, "Age of information of multi-source systems with packet management," in *Proc. IEEE Int. Commun. Conf. Workshop*, Dublin, Ireland, Jun. 2020, pp. 1–6. (cited on page 14)

[99] ——, "Whittle index-based scheduling policy for minimizing the cost of age of information," *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 54–58, Jan. 2022. (cited on page 14)

[100] ——, "Age of information of multi-user mobile edge computing systems," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1600–1614, Jul. 2023. (cited on page 15)

[101] ——, "Age of information analysis of multi-user mobile edge computing systems," in *Proc. IEEE Global Commun. Conf.*, Dec. 2021, pp. 1–6. (cited on page 15)

[102] Z. Tang, N. Yang, P. Sadeghi, and X. Zhou, "The age of information of short-packet communications: Joint or distributed encoding?" in *Proc. IEEE Int. Commun. Conf.*, Seoul, Republic of Korea, May 2022, pp. 2175–2180. (cited on pages 16, 17, and 85)

[103] Z. Tang, N. Yang, X. Zhou, and J. Lee, "Average age of information penalty of short-packet communications with packet management," in *Proc. IEEE Int. Commun. Conf.*, Rome, Italy, May. 2023, pp. 1–6. (cited on pages 16 and 17)

[104] Z. Tang, N. Yang, P. Sadeghi, and X. Zhou, "Age of information in downlink systems: Broadcast or unicast transmission?" *IEEE J. Select. Areas Commun.*, vol. 41, no. 7, pp. 2057–2070, Jul. 2023. (cited on page 18)

[105] ——, "Broadcast versus distributed short-packet transmission: An age of information perspective," in *Proc. IEEE Int. Commun. Conf.*, Rome, Italy, May. 2023, pp. 1–6. (cited on page 18)

[106] N. L. Johnson, A. W. Kemp, and S. Kotz, *Univariate discrete distributions*. John Wiley & Sons, 2005. (cited on page 26)

[107] P. Whittle, "Restless bandits: activity allocation in a changing world," *J. Appl. Probab.*, vol. 25, no. 1, pp. 287–298, Jan. 1988. (cited on pages 32 and 38)

[108] D. P. Bertsekas, *Dynamic programming and optimal control*. Belmont, MA, USA: Athena Scientific, 2000. (cited on page 33)

[109] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4268–4282, Oct. 2016. (cited on pages 44 and 46)

[110] S. Sthapit, J. Thompson, N. M. Robertson, and J. R. Hopgood, "Computational load balancing on the edge in absence of cloud and fog," *IEEE Trans. Mobile Comput.*, vol. 18, no. 7, pp. 1499–1512, Jul. 2019. (cited on page 45)

[111] C.-L. Chen, C. G. Brinton, and V. Aggarwal, "Latency minimization for mobile edge computing networks," *IEEE Trans. Mobile Comput.*, vol. 22, no. 4, pp. 2233–2247, Apr. 2023. (cited on page 51)

[112] J. Zhao, Y. Zhang, and G. Cao, "Data pouring and buffering on the road: A new data dissemination paradigm for vehicular Ad Hoc networks," *IEEE Trans. Veh. Technol.*, vol. 56, no. 6, pp. 3266–3277, Nov. 2007. (cited on page 54)

[113] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge: Cambridge University Press, 2011. (cited on page 63)

[114] M. Lopez-Benitez and F. Casadevall, "Versatile, accurate, and analytically tractable approximation for the Gaussian Q-function," *IEEE Trans. Commun.*, vol. 59, no. 4, pp. 917–922, Feb. 2011. (cited on page 68)

[115] H. Hu, K. Xiong, Y. Lu, B. Gao, P. Fan, and K. B. Letaief, "α-β AoI penalty in wireless-powered status update networks," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 474–484, 2022. (cited on pages 72, 74, and 75)

[116] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Commun. Mag.*, vol. 25, no. 3, pp. 124–130, Jun. 2018. (cited on pages 85 and 95)

[117] W. Liang, M. Zheng, J. Zhang, H. Shi, H. Yu, Y. Yang, S. Liu, W. Yang, and X. Zhao, "WIA-FA and its applications to digital factory: A wireless network solution for factory automation," *Proc. IEEE*, vol. 107, no. 6, pp. 1053–1073, Jun. 2019. (cited on page 91)

[118] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011. (cited on page 93)

[119] M. Hosseinian, J. P. Choi, S.-H. Chang, and J. Lee, "Review of 5G NTN standards development and technical challenges for satellite integration with the 5G network," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 36, no. 8, pp. 22–31, Aug. 2021. (cited on page 106)

[120] A. Goldsmith, "Adaptive modulation and coding for fading channels," in *Proc. IEEE Info. Theory Commun. Workshop*, Kruger National Park, South Africa, Jun. 1999, pp. 24–26. (cited on page 107)

[121] F. Peng, J. Zhang, and W. E. Ryan, "Adaptive modulation and coding for IEEE 802.11n," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Hong Kong, China, Mar. 2007, pp. 656–661. (cited on page 107)

[122] M. P. Mota, D. C. Araujo, F. H. Costa Neto, A. L. F. de Almeida, and F. R. Cavalcanti, "Adaptive modulation and coding based on reinforcement learning for 5G networks," in *Proc. IEEE Global Commun. Conf. Workshop*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6. (cited on page 107)