Data Freshness and End-to-End Delay in Cross-Layer Two-Tier Linear IoT Networks

Imane Cheikh [1], Essaid Sabir [1,2,*], Rachid Aouami [3], Sébastien Roy [3] and Mohamed Sadik [1]

1 NEST Research Group, LRI Lab, ENSEM, Hassan II University of Casablanca, Casablanca 20000, Morocco
2 Department of Computer Science, University of Quebec at Montreal, Montreal, QC H2L 2C4, Canada
3 Department of Electrical and Computer Engineering, University of Sherbrooke, Sherbrooke, QC J1K 2R1, Canada
* Correspondence: essaid.sabir@ieee.org

Abstract: The operational and technological structures of radio access networks have undergone tremendous changes in recent years. A displacement of priority from capacity–coverage optimization (to ensure data freshness) has emerged. Multiple radio access technology (multi-RAT) is a solution that addresses the exponential growth of traffic demands, providing degrees of freedom in meeting various performance goals, including energy efficiencies in IoT networks. The purpose of the present study was to investigate the possibility of leveraging multi-RAT to reduce each user's transmission delay while preserving the requisite quality of service (QoS) and maintaining the freshness of the received information via the age of information (AoI) metric. First, we investigated the coordination between a multi-hop network and a cellular network. Each IoT device served as an information source that generated packets (transmitting them toward the base station) and a relay (for packets generated upstream). We created a queuing system that included the network and MAC layers. We propose a framework comprised of various models and tools for forecasting network performances in terms of the end-to-end delay of ongoing flows and AoI. Finally, to highlight the benefits of our framework, we performed comprehensive simulations. In discussing these numerical results, insights regarding various aspects and metrics (parameter tuning, expected QoS, and performance) are made apparent.

Keywords: ad hoc network; age of information; cellular network; delay; IoT; multi-RATs integration; queuing theory

1. Introduction

One recent significant advancement of the information age is the Internet of Things (IoT), which provides convenient benefits, resulting in the widespread growth of mobile network services and the promotion of more comfortable and relevant lifestyles and facilities. However, this rapid development has resulted in a large rise in energy consumption, leading to greater greenhouse gas emissions and higher financial expenses for network operators. Energy costs associated with the operation of a cellular network now account for a sizable share of the global human energy footprint. As a result, network operators are searching for innovative ways to reduce and manage their energy footprints [1]. Overall, for a sensor network without energy recovery capabilities, energy-efficient communication technology is required for data transmission. A sensor will be unusable immediately after its battery is discharged (if no alternate power source is available). Therefore, it is crucial to understand and characterize the performances of sensor networks, especially in terms of delay and energy consumption. Ideally, a sensor network should have the longest operating life before requiring maintenance (such as a battery change). Consequently, it is necessary to operate such networks at the lowest possible energy consumption; this has been an ongoing area of research [2].

To address these issues, the development of advanced wireless systems and services is taking place in a heterogeneous environment where multiple RATs coexist. As a result, the complexity and cost of network deployment decrease, leading to even higher energy efficiency gains.

Currently, different radio access technologies (RATs) typically operate independently from each other. However, there is a growing demand for coordination between different RATs to meet the exponential growth in wireless traffic. Mobile users or autonomous sensor nodes can be served simultaneously by two or more RATs. Commonly, multiple connections participate in the application or transport layer, and each connection (or flow) corresponds to a single RAT (5G, LoRa, NB-IoT, LTE-M, etc.) over which the data stream [3]. The collaboration enables and maintains connectivity for universal use and provides the most appropriate services for users, regardless of time or location. With multiple radio interfaces, IoT devices are granted the ability to communicate simultaneously over different interfaces and select the "best" interface at any given moment based on a variety of parameters, such as QoS requirements, network capabilities, application properties, etc. Essentially, for each interface, there is a specific range and cost (energy, economic issues, etc.) [4].

Keeping all of these considerations in mind, the goal of this article is to address the minimization of the total delay in a multi-RAT network while taking into account data freshness. The integration of a multi-hop wireless ad hoc network and a cellular network to form a multi-RAT IoT platform constitutes the core of this paper. In such a platform, the nodes coordinate and dynamically switch between RATs, with the aim of determining the best path to the destination while ensuring the data freshness and QoS constraints are met. Furthermore, multi-hop relay technology, which is widely utilized in ad hoc networks, can also advantageously be applied to cellular networks to increase network capacity [5].

A multi-hop wireless ad hoc network consists mainly of a series of nodes communicating with each other when no centralized control and fixed infrastructure are available. Many different factors, such as the routing protocol and channel access methods, play a role in making communications possible. Wireless ad hoc networks are commonly used for commercial purposes, such as providing internet connectivity to nodes that are outside the transmission range of a wireless access point. This suggests that cellular and ad hoc networks are in many ways complementary [4]. Many studies to date have concentrated on increasing network throughput and investigating the effect of modulation order on energy efficiency. In contrast, the integration of a multi-hop wireless ad hoc network with a multi-RAT system has not been investigated on the same scale. The goal of multi-RAT optimization is to discover the collection of network components that uses the least amount of energy while maintaining network QoS criteria.

A new metric known as AoI has recently been developed to quantify the freshness of information in numerous IoT applications, such as remote monitoring applications, where information has a higher value when it is fresher [6]. From this vantage point, it appears that standard performance indicators, such as packet delay and throughput, are inadequate to accurately capture the timeliness of status information based on destination data. Blindly minimizing delay or increasing throughput, for example, may not keep status information at the destination as up-to-date as possible. Hence, relying on an explicit metric such as AoI is a proper avenue for assessing the freshness of information. It is most commonly defined as the time that has elapsed since the last status packet was received at the destination, allowing source nodes to assess the freshness of information from the destination side [7].

Aside from evaluating various queuing models and policies, we are interested in identifying and understanding alternative age optimization schemes for various queues. This research also looks at the age metric when a deadline is imposed on data packets waiting in queues, forcing their removal from the system after the deadline expires. Using a deadline that is too short results in more packets expiring, resulting in fewer status updates and a higher average age. However, a deadline that is too lengthy does not remove packets that have become very stale from the queue, resulting in wasteful usage of several resources for older packets, and eventually also a rise in the average age.

1.1. Related Work

To fully leverage multiple networks, the multi-RAT scheme has been introduced, where multiple technologies are deployed and help users deliver services appropriately. This is a promising approach that has recently received significant attention from researchers. Many publications have been devoted to the coexistence of converged and coordinated multiple RATs, in order to reduce overall network deployment complexity and costs while improving network operations maintenance requirements. Future networks are expected to support more intelligent management and integrate a range of wireless access technologies, as well as provide some degree of self-configuration, self-optimization, and self-healing [8].

Previous research in this field has mostly focused on maximizing network capacity while adhering to QoS limitations. Other research has concentrated on the resource allocation issue for parallel transmission employing several RATs [9,10]. However, the influence of delay on system performance was not included in these contributions. The fundamental issue that must be addressed is energy consumption in wireless communications. As a result, there is a rising emphasis in a range of studies on the design of energy-efficient wireless communication systems. Based on a realistic battery model, the authors of [11] present effective relay selection and energy allocation algorithms. In [12], the authors address the problem of optimal relay and RAT selection to optimize energy efficiency. Meanwhile, an energy-efficient joint radio resource management in heterogeneous multi-RAT networks is provided in [13].

Many researchers have long been interested in the capacity analysis of wireless communication networks. As far as we are aware, no current study on multi-RAT networks has examined QoS requirements in terms of throughput, reliability, end-to-end delay, and information age, while combining cellular/ad hoc network metrics and OSI model layers. Table 1 shows a comparison of the related literature and our work. Furthermore, the effect of heterogeneous networks on the age of information has not been thoroughly or appropriately investigated so far. In contrast, extensive work has been conducted on the study of ad hoc network performance metrics while taking OSI model parameters into account. In this context, the authors of [14] analyzed the end-to-end throughput behavior and stability of transmission queues in multi-hop wireless networks. Routing, random access in the MAC layer, and topology are all taken into account in their proposed model. They demonstrated that when the queues are stable, the end-to-end throughput of a given route is not impacted by a load of intermediate nodes. In [15], the authors started from the model used in [14] and studied the interaction between the PHY, MAC, and network layers. Subsequently, in [16], the authors investigated the end-to-end performance of a multi-hop wireless network for a real-time application, based on a cross-layer scheme, including the PHY, MAC, and network layers.

As mentioned in the introduction, the AoI metric has emerged as a means to assess the quality of status updates across a wireless network. According to the authors in [17], the AoI grows until a more current status update arrives at the receiver, where successful reception entails an abrupt reduction. Such a tool is applicable in applications where the maintenance of current information is crucial. Obviously enough, the time taken to propagate through the network contributes to the degree of staleness of the received updates. As a result, adequate AoI performance is achieved when status updates are provided, not only on a regular but also timely basis. The authors of [18] discuss the age minimization issue in a multi-hop network with a broad interference restriction. Among the most relevant works, authors of [19] demonstrate that the AoI may be decreased by ensuring that newer information constantly replaces older information in the transmission queue. In [20], this concept is expanded to a multi-hop scenario. The authors of [21] outline generic AoI analysis methods, then apply these AoI approaches to a variety of increasingly more complex systems, such as energy harvesting sensors broadcasting over noisy channels, parallel server systems, and queuing networks.

Many studies have been conducted on systems with time-constrained packets. The majority of them deal with the challenge of scheduling packets in order to reduce the number of packets that expire before successful transmission. When employed in the

context of a wireless sensor network (WSN), deadlines have been adopted to reduce delay and energy consumption [22]. However, few publications have investigated the deployment of deadlines from the standpoint of AoI control (e.g., [23,24]).

In conclusion, we note that the current literature on AoI has focused on many distinct types of queues, each with a particular arrival and departure procedure, queue capacity, and the number of servers.

Table 1. Comparison between our proposal and related work.

Reference	Topology	Performance Metrics	Main Objective	Relevant OSI Layers
[25]	Non-linear	E2E delay	To minimize the stringent task service delays for sensor and IoT devices, an analytical model was designed.	Network
[26]	Non-linear	Throughput, delay and Energy consumption	Provide hybrid HetNet offloading while taking into account user traffic loads by modeling the queues of each network user.	Network
[27]	Non-linear	E2E delay	In this paper, an analytical approach for determining the E2E mean response time of infrastructure network slices is proposed.	Network
[28]	Non-linear	Throughput, delay and Energy consumption	The development of a framework for analyzing efficient forwarding choices in terms of QoS parameters.	MAC/Network
Our work	Linear	E2E delay and AoI	Analyze the integration of a multi-hop wireless ad hoc network with a multi-RAT platform to optimize the energy consumption of the entire proposed system.	MAC/Network

1.2. Our Main Contributions

The core contribution of this paper centers on the elaboration of a theoretical framework for the performance evaluation of a dual-RAT or two-tier network. Tier 1 consists of an ad hoc multi-hop network relying on a short-range, low-power, and low-cost RAT (possibly in an unlicensed band) such as Zigbee. Tier 2 consists of a centralized single-hop network with a star topology and relies on a longer range RAT, such as cellular 5G, LTE-M, LoRa, etc. Although other RAT options are possible as noted, the tier 2 connections will be henceforth referred to as "cellular". All nodes are members of both networks and are equipped with both RATs. The physical topology of the network is assumed to be quasi-linear (as this corresponds to many applications of interest), with the base station or data sink at one end of the chain. More specifically, a probabilistic model is developed allowing us to jointly address the ad hoc/cellular channel properties and the cross-layer modeling. Our contribution can be summarized as follows:

- We build a complete framework to analyze the integration of a multi-hop wireless ad hoc network with a multi-RAT platform to optimize the energy consumption of the entire proposed system through delay minimization while ensuring data freshness, through the AoI metric.
- As illustrated in Figure 1, our model can be used in different environments such as tunnels, roads, bridges, etc.
- A cross-layer model is used, to replace the non-communicating layers of the OSI standard, involving synergy between network and MAC layers enabling the protocol stack to share specific information.
- We use a G/G/1 and an M/G/1 queuing model to estimate the waiting time at intermediate nodes.

- We determine the optimal average end-to-end delay and age of information. These two key QoS metrics provide interesting insights on how to define the internal parameters, thus achieving optimal performance.
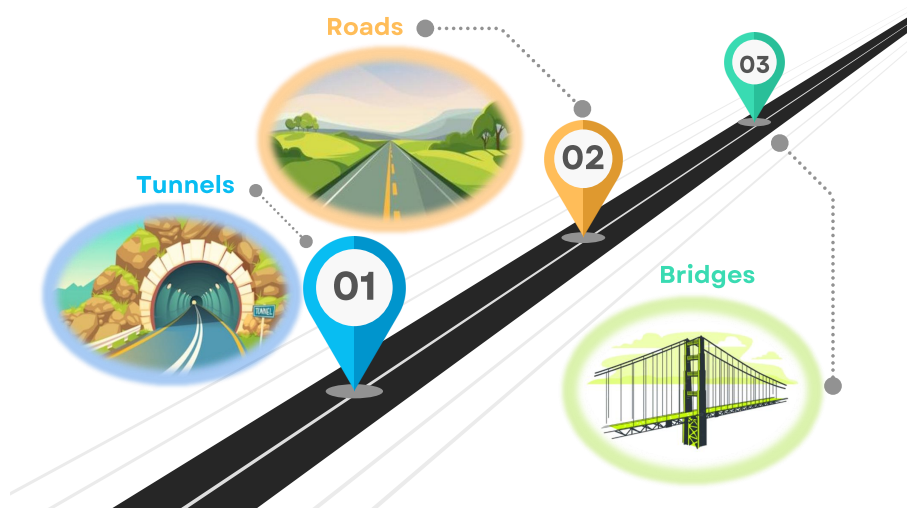

Figure 1. Use cases covered by our model.

1.3. Paper Organization

The rest of this paper is organized as follows. Problem formulation is discussed in Section 2, average delay analysis is defined in Section 3, the steady state and expressions for performance metrics are derived in Section 4, while the performance evaluation is addressed in Section 5. Finally, the concluding remarks and future works are presented in Section 6.

2. Problem Formulation

In this section, we investigate the system model, including the network topology, channel model, NET/MAC cross-layer models, energy limitations, and the proposed two-tier network incorporating both multi-hop and multi-RAT aspects.

2.1. The Setting

We consider a two-tier IoT network, including a base station and a set of $\mathcal{N} = \{1, 2, 3, \ldots, n\}$ IoT devices (such as sensors measuring temperature, pressure, vehicular speed, etc.), linearly distributed over the area, as shown in Figure 2. If the fraction of cellular traffic generated by node i is denoted ω_i, then $1 - \omega_i$ is the corresponding fraction of ad hoc traffic. At any time and for any given packet, an IoT device must choose between (i) transmission of the packet to one of its neighbors, as a stepping stone towards the final destination and (ii) sending the packet directly to the base station (cellular network). For instance, a mobile device located far from the base station and attempting to optimize its power consumption may choose to route packets through a multi-hop sequence rather than transmitting them directly to the base station. However, a device located close to the base station may receive a high number of packets to relay to the base station and thus experience a faster battery depletion. The selection strategy in a multi-RAT context can be tuned to reduce this effect, in order to equalize energy depletion across all nodes. The goal consists in optimizing and balancing energy consumption in the network while ensuring that deadlines are met and that data freshness is maintained. This is achieved through the study of two key metrics, namely end-to-end delay and AoI. The main notations and symbols included in this article are listed in Table 2.
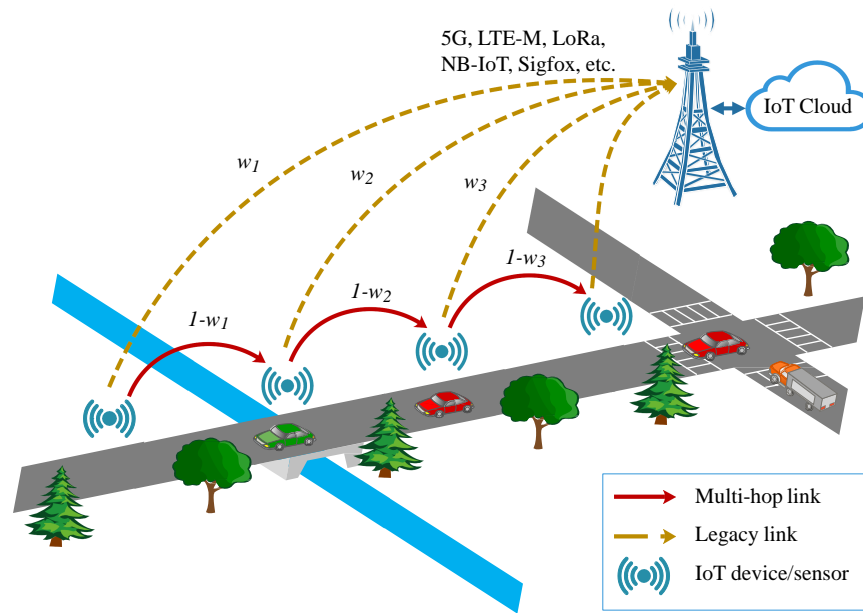
Figure 2. A two-tier IoT Network.

For clarification purposes, the model assumptions are summarized below:

1. All nodes are expected to be informed of the success or failure of their transmitted packets. In order to maintain a satisfactory level of reliability permanently, we assume that a packet is re-transmitted (if required) until success or definitive drop;
2. It is expected that each node will have two types of packets to transmit: (1) packets generated by the device itself (queue Q_i), and (2) packets received from other neighboring devices that must be forwarded until reaching the final destination (queue F_i);
3. A mobile is capable of transmitting on one interface and receiving on the other. However, it is not able to send an ad hoc and a cellular message on both network cards (no simultaneous transmissions, if we allow parallel selections, we will make multi-homing possible, (i.e., two simultaneous transmissions can be achieved)).
4. It is assumed that the system is not saturated, which entails that the F_i and/or Q_i queue might be empty at any node i.

2.2. The Channel Model

In this paper, each IoT device serves both as a relay for forwarding data generated upstream to the next node in a multi-hop chain, and as a cellular transmitter capable of reaching the base station directly. In this context, two distinct channels must be considered, i.e., (1) the ad hoc channel, and (2) the cellular channel.

2.2.1. Ad Hoc Channel

The slotted-Aloha MAC scheme is assumed for all nodes in the ad hoc network, which are also assumed to be perfectly synchronized on certain time slots. Nodes send packets using the following rule. For each time slot, each node independently tosses a coin with a certain bias p known as the Aloha medium access probability (MAP). If the result is "heads", it sends the packet in that time slot, otherwise, it does not transmit [29].

We indicate by N_i^a the average number of transmission attempts, which can be defined as:

$$N_i^a = \frac{1 - (1 - \varsigma_i)^{K_i^a}}{\varsigma_i},\tag{1}$$

where K_i^a is the maximum number of transmissions permitted by a mobile i per packet.

A transfer from i is successful if neither $i + 1$ nor any of its neighbors $\mathcal{N}(i+1)$, except i, transmits in the same time slot. The success probability ς_i for a packet at node i in ad hoc network is given by:

$$\varsigma_i = q_i \prod_{z \in \mathcal{N}(i+1) \cup (i+1) \setminus i} (1 - q_z), \tag{2}$$

where q_i indicates the attempt probability for a packet at node i.

2.2.2. Cellular Channel

A Rayleigh channel model is assumed for the cellular channel. The most essential and widely used measure of channel quality in a cellular network is the signal-to-interference-plus-noise ratio (SINR).

The SINR of the IoT device i deployed in a fixed location could be written as:

$$\gamma_i = \frac{P_i \cdot h_i}{\sigma^2 + \sum_{j \neq i} P_j \cdot h_j \omega_j}, \tag{3}$$

where P_i is the transmit power of IoT device i, h_i refers to her channel gain, which is assumed to follow a Rayleigh distribution and σ^2 is the variance of a Gaussian additive noise.

Here we look at the efficiency function $\phi(\gamma, L)$, commonly known as the packet success rate (PSR), for every user who has to send packets of L bits each to a base station is denoted as [30]:

$$\phi(\gamma_i, L) = (1 - \xi(\gamma_i))^L, \tag{4}$$

where L is the length of a given packet and $\xi(\gamma)$ is the bit error rate (BER) from one user to its serving station, which depends on the SINR used. In fact, the expression of BER varies according to the coding and modulation scheme adopted by a user. Our present study is valid for all coding and modulation schemes.

We denote by N^c the average number of transmission attempts in a cellular network, which can be expressed as follows:

$$N_i^c = \frac{1 - (1 - \phi(\gamma_i, L))^{K_i^c}}{\phi(\gamma_i, L)}, \tag{5}$$

we use K_i^c to indicate the maximum number of transmissions permitted by a mobile i per packet in a cellular network.

Table 2. Main symbols and their meanings.

Symbol	Meaning
n	Number of IoT devices
ω_i	Fraction of cellular traffic sent by node i
$1 - \omega_i$	Fraction of traffic sent over the ad hoc link by node i
N_i^a	Average number of transmission attempts in ad hoc network
N_i^c	Average number of transmission attempts in cellular network
K_i^a	Maximum number of transmissions permitted by a mobile i per packet in ad hoc network
K_i^c	Maximum number of transmissions permitted by a mobile i per packet in a cellular network
ς_i	Success probability for a packet at node i in ad hoc network
q_i	Attempt probability for a packet at node i
γ_i	SINR of device i
$\phi(\gamma_i, L)$	Efficiency function
R_i	Transmission rate (in bps)
L	Packet length (in bits)
π_i^F	Probability that queue F_i has a packet placed at the head of the line
π_i^Q	Probability that queue Q_i has a packet placed at the head of the line
f_i	Forwarding probability from queue F_i
$1 - f_i$	Forwarding probability from queue Q_i
λ_i^Q	Arrival rate in queue Q_i
λ_i^F	Arrival rate in queue F_i
t_i^c	Average packet transmission time of user i for cellular network
t_i^a	Average packet transmission time for ad hoc network of node i
W_i^F	Waiting time in queue F_i

Table 2. *Cont.*

Symbol	Meaning
W_i^Q	Waiting time in queue Q_i
B_i^F	Queuing time in queue F_i
B_i^Q	Queuing time in queue Q_i
R_i^F	Mean residual service time in queue F_i
R_i^Q	Mean residual service time in queue Q_i
$R_i^{c,F}$	Mean residual service time in queue F_i for cellular network
$R_i^{a,F}$	Mean residual service time in queue F_i for ad hoc network
$R_i^{c,Q}$	Mean residual service time in queue Q_i for cellular network
$R_i^{a,Q}$	Mean residual service time in queue Q_i for ad hoc network
V_i	Inter-arrival time of packets for a mobile device i
A_i	Age of Information for a mobile device i
A_i^F	Age of Information for a mobile device i in queue F_i
A_i^Q	Age of Information for a mobile device i in queue Q_i

2.3. Cross-Layer Architecture

Here, a cross-layer architecture is proposed, which takes into account both the network and MAC layer parameters (see Figure 3). Thus, communication and information sharing between separate layers become more efficient and flexible, and offer the possibility of global optimization.

The network layer comes first in our cross-layer architecture. It is responsible for defining the source and destination of packets and routing them through the sensor network. It manages two queues: (1) the forwarding queue F_i, and (2) the queue Q_i. Queues in the system are assumed to operate with infinite storage capacity, thus avoiding packet loss by overflow. A scheduling method such as first in first out (FIFO) is considered. In addition, a weighted fair queuing (WFQ) is used in the network layer for managing the data transmitted over each cycle. This scheme offers some flexibility and allows QoS support and packet prioritization.
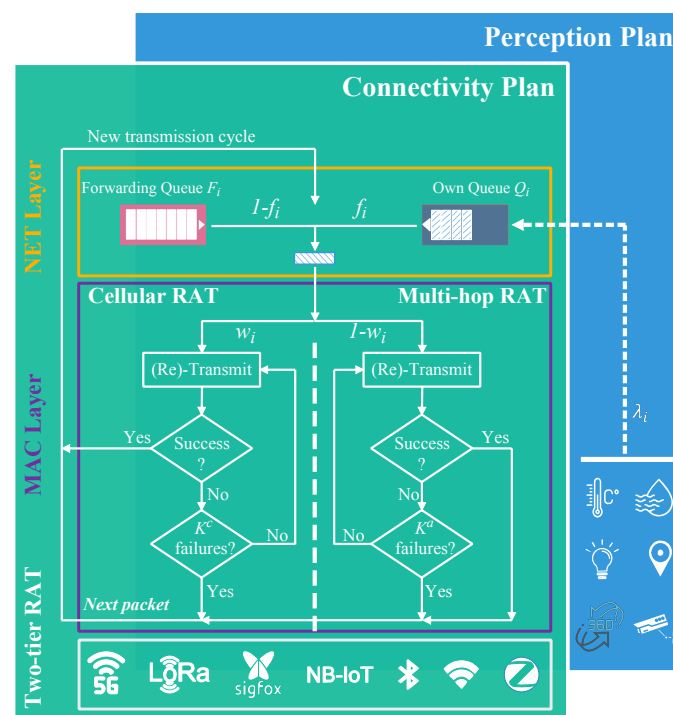

Figure 3. Two-tier IoT network packet transmission cycle and cross-layer flow chart.

2.3.1. Own Queue Q_i

This queue handles packets generated by node i itself (sensed data in the case of a sensor network), which are transmitted to their final destination (base station) through the network of neighboring sensors, or through the cellular network directly, modeled as an M/G/1 queue, where node i opts to transmit from Q_i with probability $1 - f_i$.

2.3.2. Forwarding Queue F_i

This queue contains packets from other nodes to be forwarded to the base station through one or several hops. It is modeled as a G/G/1 queue, where node i decides to forward from F_i with probability f_i. Thanks to this configuration, the nodes benefit from a certain flexibility allowing them to manage the packets transmitted by each node differently from their own packets.

The MAC layer establishes the communication media sharing rules for the different IoT devices in the network. Here, we consider a slotted-Aloha MAC protocol. Prior to any transmission attempt, a queue, either F_i or Q_i, is selected. At the beginning of each time slot, a node attempts to gain channel access with random probability q_i. Then, the head packet from the selected queue is moved from the network layer to the MAC layer where it is transmitted and retransmitted if required, until successful delivery or final drop.

2.3.3. Multi-RAT Support

As mentioned above [4], in the presence of multiple radio interfaces, IoT devices are assumed to be equipped with more than one radio interface and to select the "best" one based on multiple parameters such as user requirements, network capabilities, application properties, etc. In general, every interface has a specific range and cost (energy, economic issues, etc.). A major challenge in multi-RAT networks consists in dynamically selecting the most appropriate RAT in order to address performance goals, such as energy efficiency. Accordingly, an efficient model must be integrated at this decision stage to avoid unnecessary transfer between RATs [31].

2.3.4. Energy Limitation

A major concern in sensor network applications is the capability of operating at ultra-high energy efficiency. Nodes will shut down once their battery is discharged since there is no possibility of recharging them. Indeed, it is assumed that the nodes have no alternate power source such as harvesting, power line, etc. The deployed network must ensure that connectivity is maintained as long as possible, which raises the issue of balancing energy consumption across all nodes. If all nodes in the network consume energy at approximately the same rate, the more central nodes will remain operational and provide forwarding connectivity for a longer time. This leads to more progressive and graceful degradation of the network operation.

2.4. Routing within a Two-Tier Network

Our proposed architecture includes two tiers: (1)—the first tier is the proposed multi-hop sensor network, while (2)—the second tier consists of a cellular network.

2.4.1. Tier I: Multi-Hop Network

Sensors are presented as relay nodes, which receive/forward messages from/to their neighbors. We assume static routing, where the IoT device i forwards its packets to the mobile device $i + 1$ along a routing chain until the node responsible for relaying to the base station is reached. It is noteworthy that such a multi-hop scheme embodies many well-known benefits, in terms of QoS, generally lower transmission cost, better energy efficiency, longer device lifetime, improved spectrum efficiency/utilization, and higher self-organization capability.

2.4.2. Tier II: Cellular network

Once the packet reaches the sensor node responsible for sending the data to the base station, it is transmitted over the cellular network. We use a multi-RAT network, in which different wireless technologies are combined via separate reliable links (e.g., 5G, LoRa, Sigfox, etc.) for data transfer to the base station.

Finally, these two architectures are unified into a two-tier system to provide an efficient data transfer infrastructure in terms of delay incurred and throughput provided.

Figure 3 depicts an organizational chart that is used to fully understand the connection between the NET and MAC layers for the two-tier IoT network. It is worth mentioning that a transmission cycle comprises a number of time slots that either result in a successful transmission or a failure/drop.

3. Average Delay Analysis

Now, we focus our study on the delay, which is a performance metric corresponding to the time needed for a packet to move from source node s to the base station, by going either through the multi-hop route or directly through the cellular uplink. We first derive an expression for the entire network, then compute the delay for the cellular and multi-hop sub-systems. Finally, we estimate the arrival and departure rates of our queuing model.

Let $D_{i,j}$ be the cumulative delay that a packet experiences from the moment it is queued at node i to the moment it is transmitted over the cellular network by node j, given by:

$$D_{i,j} = \sum_{k=i}^{j} D_k^{Trans} + D_k^{Wait} + D_k^{Proc} + D_k^{Prop}. \tag{6}$$

For simplification purposes, this paper will only take into account the waiting time and the transmission time, given that both processing time D_k^{Proc} and propagation time D_k^{Prop} are negligible.

Each packet in the F_i or Q_i queue, on its way to its neighbor j, has to wait for a certain average time called waiting time (W_i^F for queue F_i and W_i^Q for Q_i). Then, in order to complete its transmission, it is directed to the second neighbor, with an ad hoc network service time t_i^a, until node j is reached, then it will be transferred to the final destination via the cellular uplink, with a service time corresponding to t_i^c.

Figure 4 illustrates the expected end-to-end delay in the entire network.

$D_{i,j}$ can be written as follows:

$$D_{i,j} = \begin{cases} \left(t_i^a + W_i^Q\right) + \left(t_j^c + W_j^F\right) + \sum_{k=i+1}^{j-1}\left(t_k^a + W_k^F\right), & j = i+1, \cdots, n. \\ W_i^Q + t_i^c, & j = i. \end{cases} \tag{7}$$
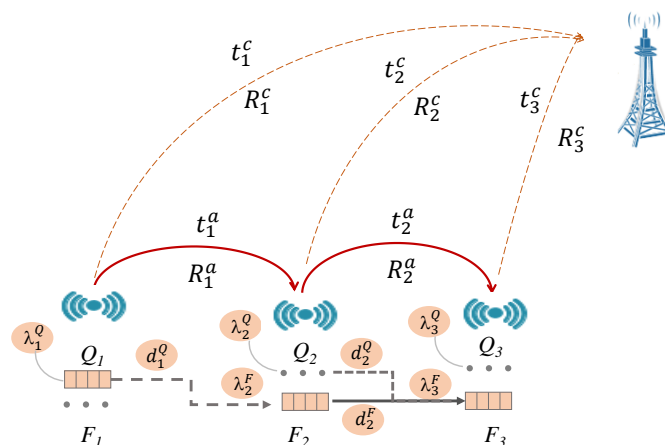

Figure 4. Expected end-to-end delay over two-tier IoT network

The delay D_i experienced at each mobile device i is obtained as:

$$D_i = \mathbb{E}_j[D_{i,j}] = \sum_{j=i}^{n} D_{i,j}\varphi(i,j), \tag{8}$$

where $\varphi(i,j)$ is the probability of sending packets over the multi-hop network from node i to node j, where the latter will forward the packet to the base station via the cellular uplink, given by:

$$\varphi(i,j) = \begin{cases} \omega_j \prod_{k=i}^{j-1}(1-\omega_k), & j < n, \\ \prod_{k=i}^{j-1}(1-\omega_k), & j = n, \end{cases} \tag{9}$$

where ω_j denotes the fraction of cellular traffic sent by node j.

The average delay generated is obtained as follows:

$$D = \sum_i D_i \phi_i, \tag{10}$$

where ϕ_i is the fraction of the total load contributed by node i, expressed as follows:

$$\phi_i = \frac{\pi_i^F + \pi_i^Q}{\sum_i \pi_i^F + \pi_i^Q}. \tag{11}$$

Next, we will determine t_i^c, t_i^a, W_i^F and W_i^Q.

3.1. Delay over Cellular Sub-System

Our heterogeneous environment makes multi-RAT systems suitable, where each RAT operates independently from others. Given the possible radio technologies for the tier 2 subsystem, some (e.g., 5G, 6G, NB-IoT,...) are characterized by a deterministic multiple access channel, while others (e.g., LoRa, Sigfox, ...) rely on contention to gain access to a shared medium.

The random variable t_i^c corresponds to the average packet transmission time of user i for tier 2, given by:

$$t_i^c = \begin{cases} N_i^c \frac{L_i}{R_i}, & \text{Deterministic multiple access (5G, 6G, NB-IoT,\dots),} \\ \frac{N_i^c}{\phi(\gamma_i,L)}, & \text{Random access through contention (LoRa, Sigfox, \dots),} \end{cases} \tag{12}$$

where incoming packets are transmitted by user i at a rate R_i (in bps).

We use π_i^F (resp. π_i^Q) to indicate the probability that queue F_i (resp. Q_i) has a packet ready to be transmitted. Moreover, let $\pi_{i,s}^F$ be the probability that queue F_i has a packet ready to be forwarded. Thus, we have:

$$\pi_i^F = \sum_{s=1}^{i-1} \pi_{i,s}^F. \tag{13}$$

3.2. Delay over Multi-Hop Sub-System

We use t_i^a to represent the average packet transmission time for the ad hoc network (tier 1) at node i, given by:

$$t_i^a = \frac{N_i^a}{\varsigma_i}. \tag{14}$$

3.3. Waiting Time

The waiting time in queue F_i (resp. Q_i) is composed of two elements: (1) the queuing time B_i^F (resp. B_i^Q); and (2) the mean residual service time R_i^F (resp. R_i^Q). The latter is divided into two terms: (1) the mean residual service time of a tier 2 packet in service R_i^c; (2) the mean residual service time of an ad hoc (tier 1) packet in service R_i^a.

The average waiting time at node i in queue F_i (resp. Q_i) is defined as:

$$W_i^F = R_i^F + B_i^F, \tag{15}$$

$$W_i^Q = R_i^Q + B_i^Q. \tag{16}$$

3.3.1. Mean residual service time at node i:

Any arriving packet should wait until the packet in service is delivered. The latter can be a packet from the F_i queue or a packet from the Q_i queue, destined directly to the base station (tier 2 network) or next neighbor (j) (ad hoc/tier 1 network). The average residual service time observed by a given packet in F_i or Q_i is denoted:

$$R_i^F = \omega_i R_i^{c,F} + (1 - \omega_i) R_i^{a,F}, \tag{17}$$

$$R_i^Q = \omega_i R_i^{c,Q} + (1 - \omega_i) R_i^{a,Q}. \tag{18}$$

Leveraging renewal theory and the method presented in [16], it can be shown that the mean residual service time in F_i for cellular network $R_i^{c,F}$ and the mean residual service time in F_i for ad hoc network $R_i^{a,F}$ (resp. $R_i^{c,Q}$ and $R_i^{a,Q}$) can be expressed as follows:

$$\textbf{Queue F:} \begin{cases} R_i^{c,F} = \dfrac{t_i^{c(2)}}{2t_i^c} + \dfrac{1}{2}, \\[3mm] R_i^{a,F} = \dfrac{t_i^{a(2)}}{2t_i^a} + \dfrac{1}{2}, \end{cases} \tag{19}$$

$$\textbf{Queue Q:} \begin{cases} R_i^{c,Q} = \dfrac{t_i^{c(2)}}{2t_i^c} + \dfrac{1}{2}, \\[3mm] R_i^{a,Q} = \dfrac{t_i^{a(2)}}{2t_i^a} + \dfrac{1}{2}, \end{cases} \tag{20}$$

where $t_i^{a(2)}$ and $t_i^{c(2)}$ designate the second moment of the service time for the ad hoc and cellular network, respectively, given by [4]:

$$t_i^{a(2)} = \frac{N_i^{a(2)} + N_i^a(1 - \varsigma_i)}{\varsigma_i^2}, \tag{21}$$

$$t_i^{c(2)} = \begin{cases} N_i^{c(2)} \dfrac{L_i}{R_i}, & \text{Deterministic multiple access (5G, 6G, NB-IoT,\dots),} \\[3mm] \dfrac{N_i^{c(2)} + N_i^c(1 - \phi(\gamma_i, L))}{\phi(\gamma_i, L)^2}, & \text{Random access through contention (LoRa, Sigfox, \dots),} \end{cases} \tag{22}$$

where:

$$N_i^{a(2)} = N_i^a + \frac{2(1 - \varsigma_i)}{\varsigma_i^2} - \frac{2(1 - \varsigma_i)^{K_i^a}(K_i^a - (1 - \varsigma_i)(K_i^a - 1))}{\varsigma_i^2}, \tag{23}$$

$$N_i^{c(2)} = N_i^c + \frac{2(1 - \phi(\gamma_i, L))}{\phi(\gamma_i, L)^2} - \frac{2(1 - \phi(\gamma_i, L))^{K_i^c}(K_i^c - (1 - \phi(\gamma_i, L))(K_i^c - 1))}{\phi(\gamma_i, L)^2}. \tag{24}$$

3.3.2. Queuing time at node i:

Once a packet enters the forwarding queue (resp. its own queue), it must wait until the available remaining packets are served before being processed. Once at the head of the forwarding queue (resp. its own queue), it must wait for the packets that will be served before it from its own queue m_i^Q (resp. the forwarding queue m_i^F). The queuing time in forwarding queue B_i^F (resp. its own queue B_i^Q) can, therefore, be written as:

$$B_i^F = m_i^Q \left(1 + m_i^F \right) ((1 - \omega_i)t_i^a + \omega_i t_i^c), \tag{25}$$

$$B_i^Q = m_i^F \left(1 + m_i^Q \right) ((1 - \omega_i)t_i^a + \omega_i t_i^c), \tag{26}$$

where m_i^F is the number of previously entered packets waiting in the forwarding queue. A packet at the top of the forwarding queue (resp. its own queue) ready for transmission must wait a certain number of cycles X (random variable) before it can move to the MAC layer. X corresponds to the number of cycles required to serve packets from Q_i (resp. from F_i). The probability of waiting k cycles is $P[X = k] = (1 - f_i)^k f_i$. The expected value of random variable X is: $E[X] \simeq m_i^Q \simeq \frac{1-f_i}{f_i}$ (resp. $m_i^F \simeq \frac{1}{m_i^Q} \simeq \frac{f_i}{1-f_i}$).

Based on Little's formula $m_i^F = \lambda_{i,s}^F W_i^F$ for queue F_i ($m_i^Q = \lambda_{i,s}^Q W_i^Q$ for queue Q_i). The waiting time at node i in queue F_i (resp. Q_i) is obtained by using Equation (15) (resp. (16)) and (25) (resp. (26)) as specified below:

$$W_i^F = \frac{R_i^F + ((1 - \omega_i)t_i^a + \omega_i t_i^c)\left(\frac{1-f_i}{f_i}\right)}{1 - \lambda_i^F((1 - \omega_i)t_i^a + \omega_i t_i^c)\left(\frac{1-f_i}{f_i}\right)}, \tag{27}$$

$$W_i^Q = \frac{R_i^Q + ((1 - \omega_i)t_i^a + \omega_i t_i^c)\left(\frac{f_i}{1-f_i}\right)}{1 - \lambda_i^Q((1 - \omega_i)t_i^a + \omega_i t_i^c)\left(\frac{f_i}{1-f_i}\right)}. \tag{28}$$

3.4. Outer Flow

This performance metric measures the rate (per time slot) at which packets are withdrawn from the queues after either a successful transmission or a drop. We identify two independent departure flows: (1) The first flow comprised of packets removed from queue Q_i, and denoted

$$d_i^Q = (1 - \omega_i)\frac{1 - \pi_i^Q f_i}{t_i^a} + \omega_i\frac{1 - \pi_i^Q f_i}{t_i^c}, \tag{29}$$

and (2) The second flow comprised of packets removed from queue F_i, given by:

$$d_{i,s}^F = \begin{cases} \frac{\pi_{i,s}^F f_i}{t_i^a}, & \text{ad hoc link,} \\ \\ \frac{\pi_{i,s}^F f_i}{t_i^c}, & \text{Cellular link.} \end{cases} \tag{30}$$

The departure rate d_i^F experienced at each mobile device i is expressed as

$$d_i^F = \mathbb{E}_s[d_{i,s}^F] = \sum_s \frac{\pi_{i,s}^F f_i}{t_i^a}(1 - \omega_i) + \frac{\pi_{i,s}^F f_i}{t_i^c}\omega_i. \tag{31}$$

3.5. Inner Flow

The inner flow is defined as the rate per time slot at which packets arrive at the queues. We identify two independent arrival flows, with (1) the first flow being composed of packets generated by IoT device i. Let us assume that packets destined for user i will be served with a Poisson distribution whose parameter λ_i^Q corresponds to the average packet arrival

rate in its own queue Q_i, where any packet is composed of L bits. The resulting source rate (in bits/second) is therefore indicated by $L\lambda_i^Q$. The second flow (2) is composed of packets from another neighbor, transmitted via the multi-hop channel. Here, IoT device i acts as a cooperative relay to transmit data packets to node j responsible for transferring data to the base station.

$\lambda_{i,s}^F$ denotes the average packet arrival rate in forwarding queue F_i of node i from a source s, and is expressed as:

$$\lambda_{i,s}^F = \begin{cases} 0 & i = s, \\ d_s^Q(1 - \pi_s^Q f_s)\prod_{z=s}^{i-1}(1 - \omega_z)\left(1 - (1 - \varsigma_z)^{K_z^a}\right). & \forall i \in \mathcal{N}, \ \forall s = 1, 2, \cdots, i-1. \end{cases} \tag{32}$$

Proof. Here we sketch a simplified proof using events decomposition. Let us consider events A and B as follows:

- **Event A:** Traffic generated by mobile device i has departed from queue Q_s;
- **Event B:** All transmissions over successive hops from mobile device s to node i have been successfully achieved.

We can easily verify that:

$P(A) = d_s^Q(1 - \pi_s^Q f_s); P(B) = \prod_{z=s}^{i-1}(1 - \omega_z)(1 - (1 - \varsigma_z)^{K_z^a}).$

As a result, the arrival rate can be expressed as

$$\lambda_{i,s}^F = P(A \cap B), \tag{33}$$

which completes the proof. The total arrival rate at node i is then given by:

$$\lambda_i^F = \sum_s \lambda_{i,s}^F. \tag{34}$$

$\square$

4. Steady State

We estimate in this section the performance metrics in terms of throughput, delay and AoI under steady-state conditions.

In the steady state, the long-term arrival rate is equal to the long-term departure rate. This corresponds to the rate balance Equation (RBE). Thus the F_i (resp. Q_i) queue is stable if its departure rate is at least equal to its arrival rate.

It is written:

$$\begin{cases} \lambda_i^F = d_i^F, \\ \lambda_i^Q = d_i^Q. \end{cases} \forall i \in \mathcal{N}. \tag{35}$$

Indeed, given that we have defined the last two metrics, it is possible to determine the expression of the average load $\pi_{i,s}^F$ (resp. π_i^Q) at each mobile device i and for each queue. The RBE results in a linear system, where the queuing system load of F_i, denoted $\boldsymbol{\pi}^F = (\pi_{1,s}^F, \pi_{2,s}^F, \cdots, \pi_{i,s}^F)$, is given by:

$$\boldsymbol{\pi}^F = \boldsymbol{G}^{-1} \cdot \boldsymbol{A}, \tag{36}$$

where $\boldsymbol{G}$ is an $I \times I$ matrix and $\boldsymbol{A}$ is a column vector with the dimensionality $I \times 1$.

Proof. Consider the following term obtained by using Equations (31), (34), and (35):

$$\alpha_{i,s} = \frac{d_s^Q(1 - \pi_s^Q f_s)\prod_{z=s}^{i-1}(1 - \omega_z)\left(1 - (1 - \varsigma_z)^{K_z^a}\right)}{f_i\left(\frac{(1-\omega_i)}{t_i^a} + \frac{\omega_i}{t_i^c}\right)}, \quad \forall i \in \mathcal{N}, \ \forall s = 1, 2, \cdots, i-1. \tag{37}$$

Then:

$$
\overbrace{\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}}^{G} \begin{pmatrix} \pi^F_{1,1} \\ \pi^F_{2,1} \\ \pi^F_{2,2} \\ \pi^F_{3,1} \\ \pi^F_{3,2} \\ \pi^F_{3,3} \\ \vdots \end{pmatrix} = \overbrace{\begin{pmatrix} 0 \\ \alpha_{2,1} \\ 0 \\ \alpha_{3,1} \\ \alpha_{3,2} \\ 0 \\ \vdots \end{pmatrix}}^{A},
$$
(38)

we obtain:

$$
\pi^F = G^{-1} \cdot A,
$$
(39)

The queuing system load of Q_i noted $\pi^Q = (\pi^Q_1, \pi^Q_2, \cdots, \pi^Q_i)$ and given by:

$$
\pi^Q = O^{-1} \cdot Y,
$$
(40)

where O is a $I \times I$ matrix and Y is a column vector with dimensionality $I \times 1$. $\square$

Proof. Consider the following term obtained by using Equations (29) and (35):

$$
\beta_i = \frac{1}{f_i} \left(1 - \frac{\lambda^Q_i}{\frac{(1-\omega_i)}{t^a_i} + \frac{\omega_i}{t^c_i}} \right).
$$
(41)

Then:

$$
\overbrace{\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}}^{O} \cdot \begin{pmatrix} \pi^Q_1 \\ \pi^Q_2 \\ \pi^Q_3 \\ \pi^Q_4 \\ \pi^Q_5 \\ \pi^Q_6 \\ \vdots \end{pmatrix} = \overbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \vdots \end{pmatrix}}^{Y},
$$
(42)

we obtain:

$$
\pi^Q = O^{-1} \cdot Y.
$$
(43)

$\square$

Age of Information

We are interested in applications where the objective is to continuously communicate the most recently updated state of a time-varying process to a given monitor. As an example, a device sends packets containing a certain state (e.g., sensor data, a list of neighboring nodes) to a network manager on a regular basis to keep the state tracked by the network manager relatively fresh at all times. IoT devices attempt to report their status to the receiver side as soon as possible. The recently proposed AoI metric measures the timeliness and freshness of status updates from various IoT devices at the destination node. It is assumed that time is divided into equal-length slots, and each status update packet is transmitted using exactly one time slot.

In Figure 5, we show the evolution of AoI $A_i(t)$ over time where A_k indicates the *kth* peak age, dropping points correspond to the instants when an update packet is received, resulting in a lower age value (i.e., the current time minus the generation time of the new update packet). We can observe from this figure that the AoI is linearly increasing over time and decreasing in case the packets are successfully received.
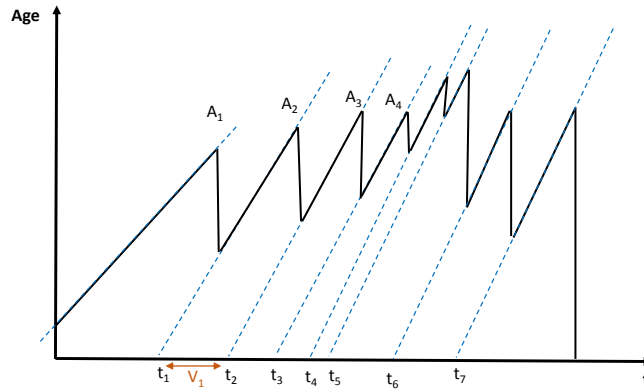
Figure 5. A sample of the evolution of AoI over time.

Given $A_i(t)$, the average age of device i can be defined as follows:

$$A_{ave} = \lim_{T \to \infty} \frac{1}{T} \int_{t=0}^{T} A_i(t) dt. \tag{44}$$

Nevertheless, the AoI metric is difficult to analyze. Furthermore, in many systems, it is often the peak state information delay that determines the performance loss. Accordingly, we focus instead on the average peak state age (PAoI) representing the maximum age of the information before a new update is received. As defined in [32], and for a given queuing model, the generalization of PAoI is given by:

$$A_i = \mathbb{E}[V_i + W_i + t_i], \ \forall i \in \mathcal{N}, \tag{45}$$

where V_i denotes the inter-arrival time of packets for mobile device i, given by:

$$V_i = V_{i,s}^F + V_i^Q = \frac{1}{\lambda_{i,s}^F} + \frac{1}{\lambda_i^Q}. \tag{46}$$

For our model, the peak age of information (PAoI) for a packet in the F_i (resp. Q_i) queue from source node s to a given neighbor i is given by:

$$A_{i,s}^F = \mathbb{E}\left[V_s^Q + W_s^Q + (1 - \omega_s)t_s^a + \sum_{j=s+1}^{i-1} \left(V_{j,s}^F + W_j^F + (1 - \omega_j)t_j^a \right) + V_{i,s}^F + W_i^F + ((1 - \omega_i)t_i^a + \omega_i t_i^c) \right], \ \forall i \in \mathcal{N}, \tag{47}$$

$$A_i^Q = \mathbb{E}\left[V_i^Q + W_i^Q + ((1 - \omega_i)t_i^a + \omega_i t_i^c) \right], \ \forall i \in \mathcal{N}, \tag{48}$$

The PAoI obtained at each mobile device i is expressed as follows:

$$A_i = \underbrace{f_i \sum_{s=1}^{i-1} A_{i,s}^F}_{\text{Packets received from neighbors}} + \overbrace{(1 - f_i) A_i^Q}^{\text{Own packets}}. \tag{49}$$

5. Performance Evaluation

This section examines the behavior of the end-to-end delay and the age of information when the fundamental parameters (f_i, q_i, ω_i, λ_i^Q) change. For illustrative purposes, a network of four sensor nodes ($n = 4$) and one base station is considered.

The simulation was carried out under three different scenarios, using MathWorks Matlab R2022b:

1. **Setting 1:** For this first instance, we assumed that the first three nodes ($i = 1, 2, 3$) have the same fraction of cellular traffic ($\omega_i = 0.5$), and the node closest to the base station ($i = 4$) needs to relay received packets to the base station, hence we would always retain $\omega_4 = 1$ for all subsequent cases.

2. **Setting 2:** In the second scenario, the node farthest from the base station ($i = 1$) sends all of its packets directly to the base station ($\omega_1 = 1$), the second node ($i = 2$) sends 75% ($\omega_2 = 0.75$) of its packets to the base station and the rest (25% of its packets) to the neighboring node, and third node ($i = 3$) sends 50% ($\omega_3 = 0.5$) of its packets directly to the base station and the rest to the neighboring node. Finally, the last node ($i = 4$) always delivers data with ($\omega_4 = 1$) to the base station.

3. **Setting 3:** For the final case, consider that the node closest to the base station ($i = 1$) sends 25% of its packets to the base station ($\omega_1 = 0.25$), the second node ($i = 2$) sends 50% ($\omega_2 = 0.5$) of its packets to the base station and the rest to the neighboring node, and the third node ($i = 3$) sends 75% ($\omega_3 = 0.75$) of its packets to the base station and the rest to the neighboring node. Finally, the last node ($i = 4$) always provides data to the base station with ($\omega_4 = 1$).

It is worth noting that $f_1 = 0$ as IoT sensor 1 has no predecessor sensor.

5.1. Packet Delay

5.1.1. Forwarding Probability f_i

Figures 6–8 depict the delay experienced by each mobile device as the forwarding probability changes with bit error rate in the first, second, and third cases, respectively. We can clearly see that using a bad channel ((a), $\xi(\gamma_i) = 10^{-1}$) implies a very high delay value, progressing to a lower delay value for a fair channel ((b), $\xi(\gamma_i) = 10^{-2}$), and finally reaching a minimal delay by using a good channel ((c),$\xi(\gamma_i) = 10^{-6}$). We also notice that the first node does not experience any change in the delay value (a stable delay) because it always has a zero forwarding probability, whereas for the other nodes, the closer we are to the base station, the greater the delay, and an increase in forwarding probability has a direct influence on the obtained latency. This is to be expected, since transmission to neighboring nodes (through the ad hoc network) is preferred, thus nodes closest to the base station will obtain more data packets to transfer. For the third scenario, we notice that node 3 has the longest delay for a good channel, which can be explained by the fact that nodes 1 and 2 transfer more packets to the latter since $w_1 = 0.25$ and $w_2 = 0.5$, and also by the fact that node 3 uses the cellular network ($w_3 = 0.75$) for the transmission of most of the packets received, which further increases the delay.
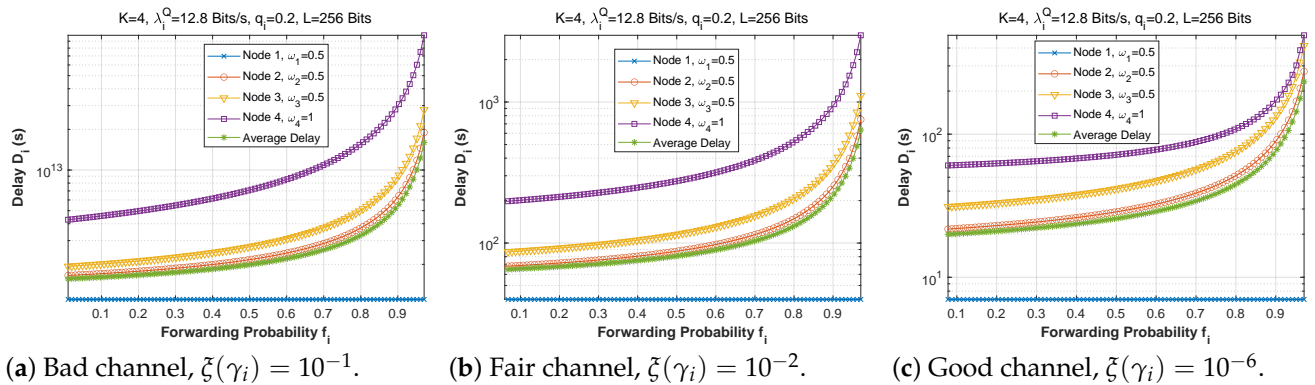

(**a**) Bad channel, $\xi(\gamma_i) = 10^{-1}$. (**b**) Fair channel, $\xi(\gamma_i) = 10^{-2}$. (**c**) Good channel, $\xi(\gamma_i) = 10^{-6}$.

Figure 6. Setting 1: The delay experienced at each mobile device when varying the forwarding probability with the bit error rate $\xi(\gamma_i)$.

(**a**) Bad channel, $\xi(\gamma_i) = 10^{-1}$.　(**b**) Fair channel, $\xi(\gamma_i) = 10^{-2}$.　(**c**) Good channel, $\xi(\gamma_i) = 10^{-6}$.
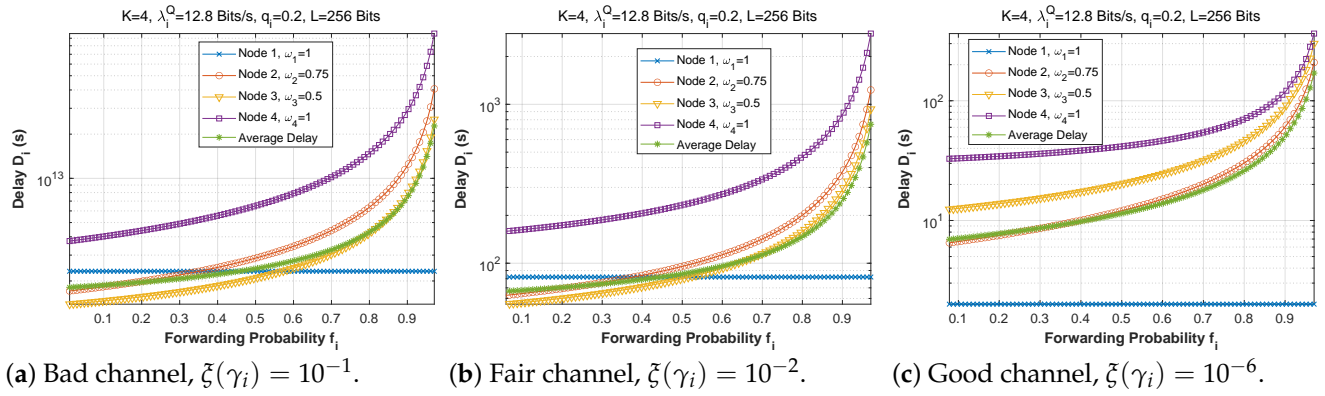
Figure 7. Setting 2: The delay experienced at each mobile device when varying the forwarding probability with the bit error rate $\xi(\gamma_i)$.

(**a**) Bad channel, $\xi(\gamma_i) = 10^{-1}$.　(**b**) Fair channel, $\xi(\gamma_i) = 10^{-2}$.　(**c**) Good channel, $\xi(\gamma_i) = 10^{-6}$.
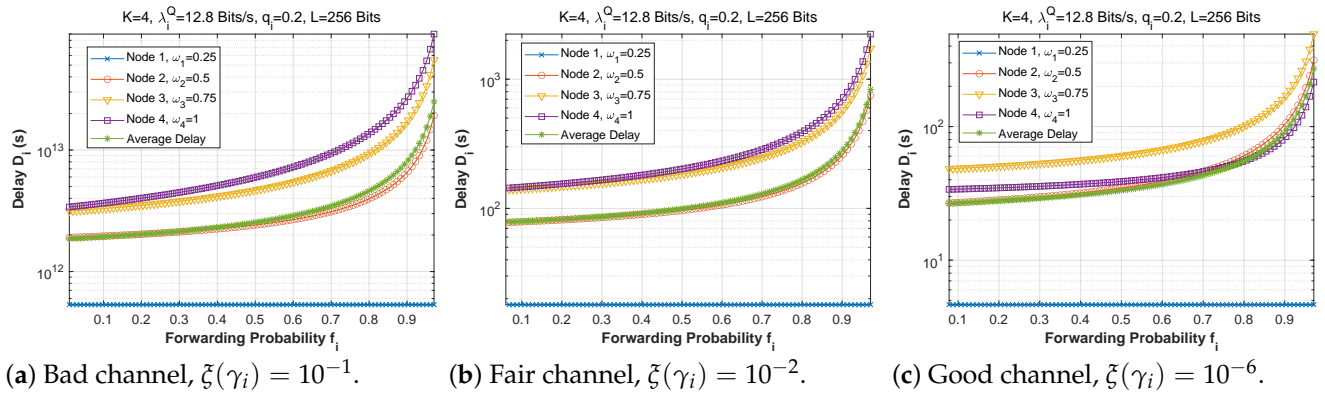
Figure 8. Setting 3: The delay experienced at each mobile device when varying the forwarding probability with the bit error rate $\xi(\gamma_i)$.

Figures 9–11 demonstrate the delay encountered by each mobile device as a function of the forwarding probability when the arrival rate in the queue varies for the first, second, and third scenarios. We notice that for heavy traffic in scenario 2 (Figure 10a, $\lambda_i^Q = 1280$ bits/s), the delay rises slowly with the increase of the forwarding probability until a maximum is reached when the forwarding probability is close to 1. However, for moderate (Figure 10b, $\lambda_i^Q = 128$ bits/s) and low traffic (Figure 10c, $\lambda_i^Q = 12.8$ bits/s), the delay starts to rise sharply for $f_i > 0.7$. Furthermore, while the behavior is approximately the same for Figure 10b,c, the curves are slightly lower in the low-traffic case, and in both cases are considerably lower than in Figure 10a. Indeed, as the traffic is reduced, the waiting time in the forwarding queue is likewise reduced. It is noteworthy that in scenarios 1 and 3, the delay curves are nearly the same shape for all three traffic levels, with very slight improvement at reduced traffic.
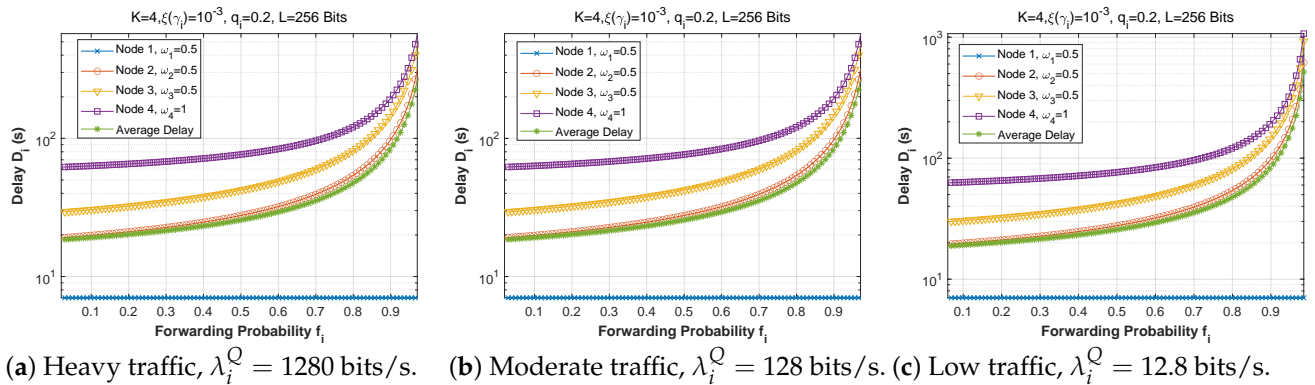

(**a**) Heavy traffic, $\lambda_i^Q = 1280$ bits/s.　(**b**) Moderate traffic, $\lambda_i^Q = 128$ bits/s.　(**c**) Low traffic, $\lambda_i^Q = 12.8$ bits/s.

Figure 9. Setting 1: The delay experienced at each mobile device when varying the forwarding probability with the arrival rate λ_i^Q.

(**a**) Heavy traffic, $\lambda_i^Q = 1280$ bits/s. (**b**) Moderate traffic, $\lambda_i^Q = 128$ bits/s. (**c**) Low traffic, $\lambda_i^Q = 12.8$ bits/s.
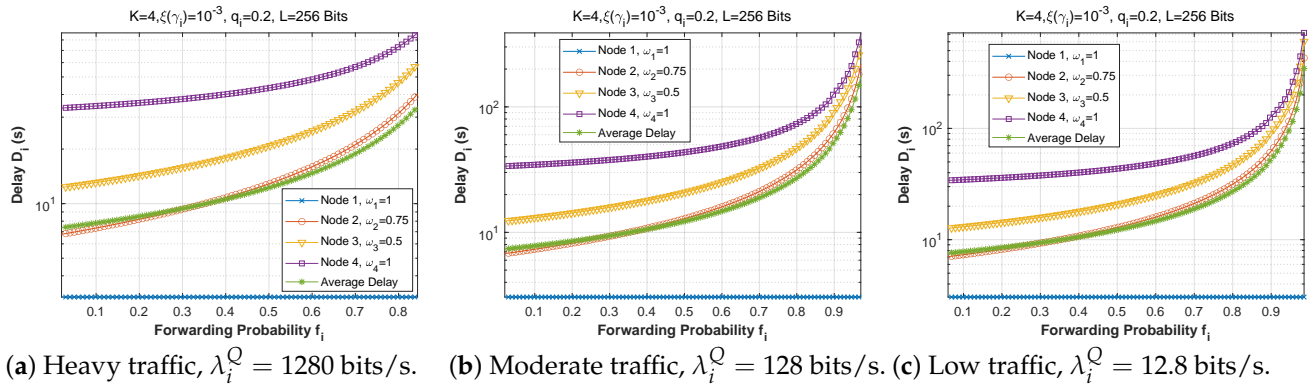
Figure 10. Setting 2: The delay experienced at each mobile device when varying the forwarding probability with the arrival rate λ_i^Q.

(**a**) Heavy traffic, $\lambda_i^Q = 1280$ bits/s. (**b**) Moderate traffic, $\lambda_i^Q = 128$ bits/s. (**c**) Low traffic, $\lambda_i^Q = 12.8$ bits/s.

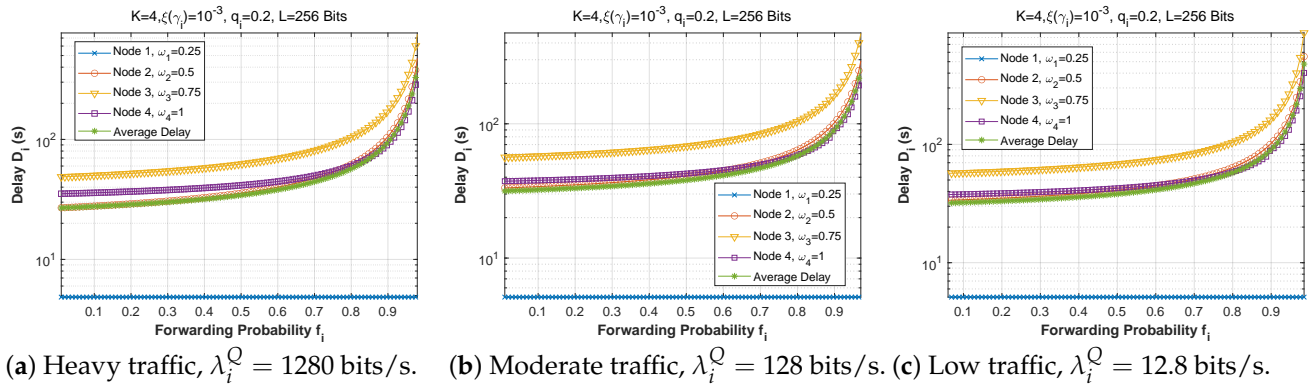**Figure 11.** Setting 3: The delay experienced at each mobile device when varying the forwarding probability with the arrival rate λ_i^Q.

5.1.2. Attempt Probability q_i

Figures 12–14 depict how latency varies as a function of attempt probability when the bit error rate changes in the first, second, and third scenarios, respectively. In the first and second scenarios with a bad channel ((a), $\xi(\gamma_i) = 10^{-1}$), we can observe that the delay is unstable, reaching very high values of up to 6×10^{12} s. However, in scenario 3, the delay is more steady, reaching a minimum when the attempt probability is between 0.2 and 0.5.

For the fair ((b), $\xi(\gamma_i) = 10^{-2}$) and good ((c), $\xi(\gamma_i) = 10^{-6}$) channels, we can see that for very small values of the attempt probability, the system is unstable. Then, for an attempt probability between 0.1 and 0.5, we have a minimal delay. For a value greater than 0.5, the delay begins to increase, which is quite normal given that the system relies heavily on the ad hoc network. We also notice that the node closest to the base station suffers a greater transmission delay than the other nodes, which can be explained by the fact that it receives several packets to transmit, which results in congestion of the queue, and therefore an increase in the transmission delay. For scenario 2, node 1 delivers all of its packets directly to the base station ($w_1 = 0$), which explains its stable latency irrespective of the value of q_i. In scenario 3, node 3 has the longest delay since it obtains more packets from neighboring nodes ($w_1 = 0.25$ and $w_2 = 0.5$), but node 1 sends most of its packets across the ad hoc network ($w_1 = 0.25$), thus, the change in q_i has a significant effect on its transmission delay.
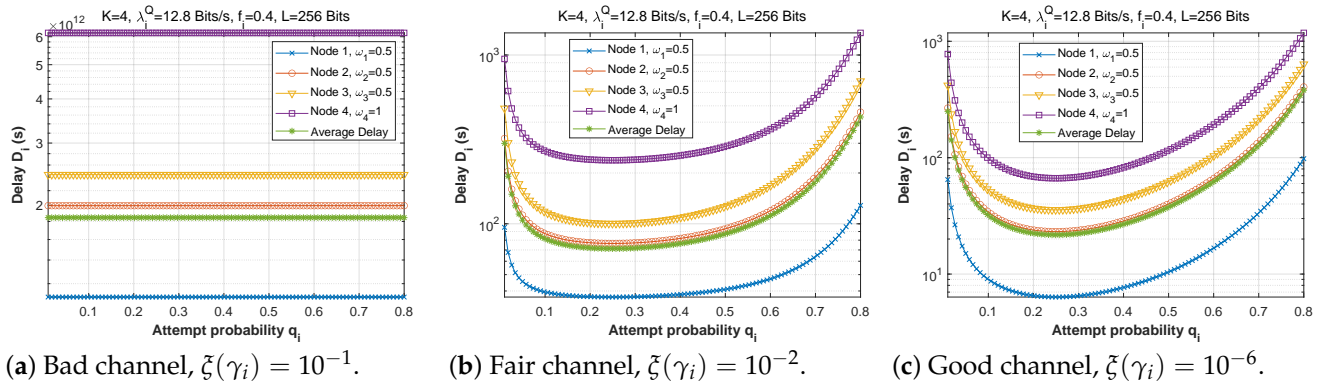
(**a**) Bad channel, $\xi(\gamma_i) = 10^{-1}$.   (**b**) Fair channel, $\xi(\gamma_i) = 10^{-2}$.   (**c**) Good channel, $\xi(\gamma_i) = 10^{-6}$.

Figure 12. Setting 1: The delay experienced at each mobile device when varying the attempt probability with the bit error rate $\xi(\gamma_i)$.
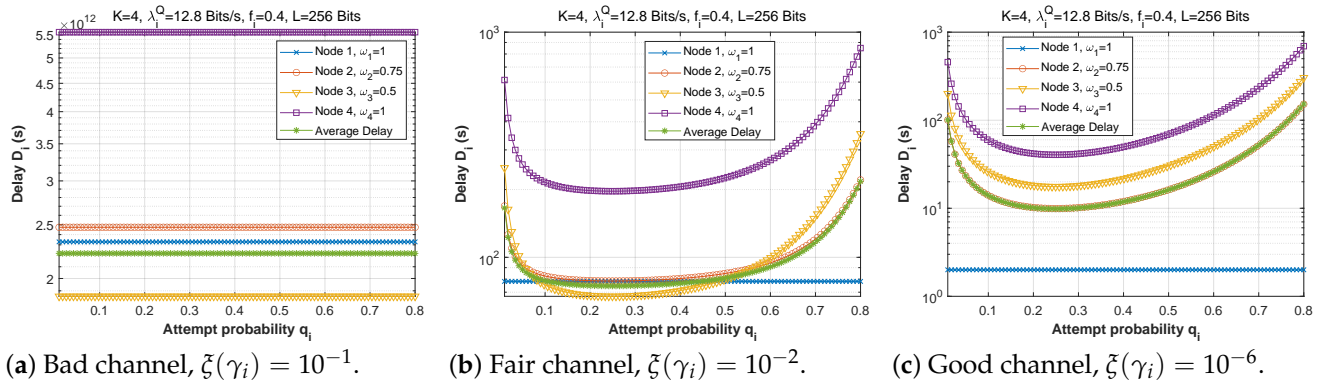

(**a**) Bad channel, $\xi(\gamma_i) = 10^{-1}$.   (**b**) Fair channel, $\xi(\gamma_i) = 10^{-2}$.   (**c**) Good channel, $\xi(\gamma_i) = 10^{-6}$.

Figure 13. Setting 2: The delay experienced at each mobile device when varying the attempt probability with the bit error rate $\xi(\gamma_i)$.
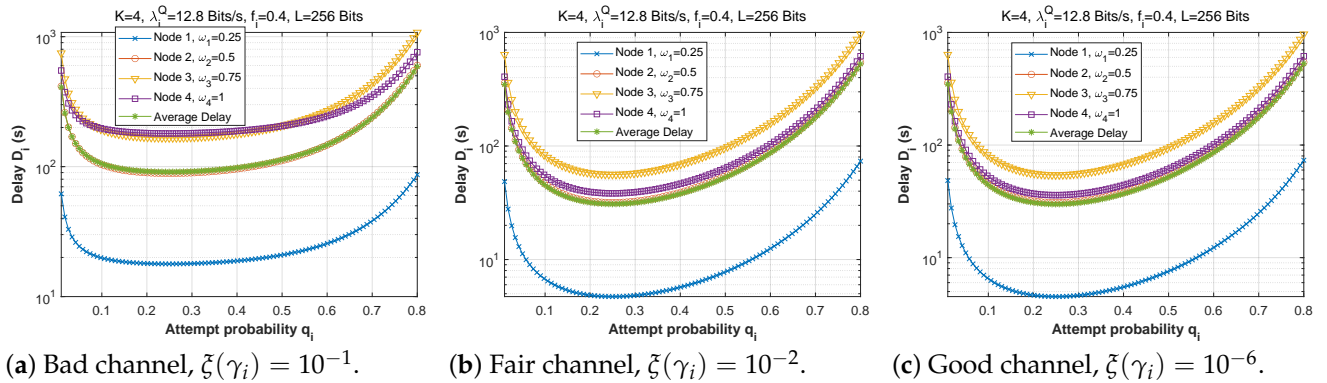

(**a**) Bad channel, $\xi(\gamma_i) = 10^{-1}$.   (**b**) Fair channel, $\xi(\gamma_i) = 10^{-2}$.   (**c**) Good channel, $\xi(\gamma_i) = 10^{-6}$.

Figure 14. Setting 3: The delay experienced at each mobile device when varying the attempt probability with the bit error rate $\xi(\gamma_i)$.

Figure 15 demonstrates the delay encountered by each mobile device as a function of the attempt probability when the arrival rate in the queue is varied (heavy traffic, $\lambda_i^Q = 1280$ bits/s, moderate traffic, $\lambda_i^Q = 128$ bits/s, low traffic, $\lambda_i^Q = 12.8$ bits/s) for the first, second, and third scenarios. We conclude that the fluctuation of q_i is unaffected by traffic density since the transmission delay is the same for all traffic categories in all three scenarios.
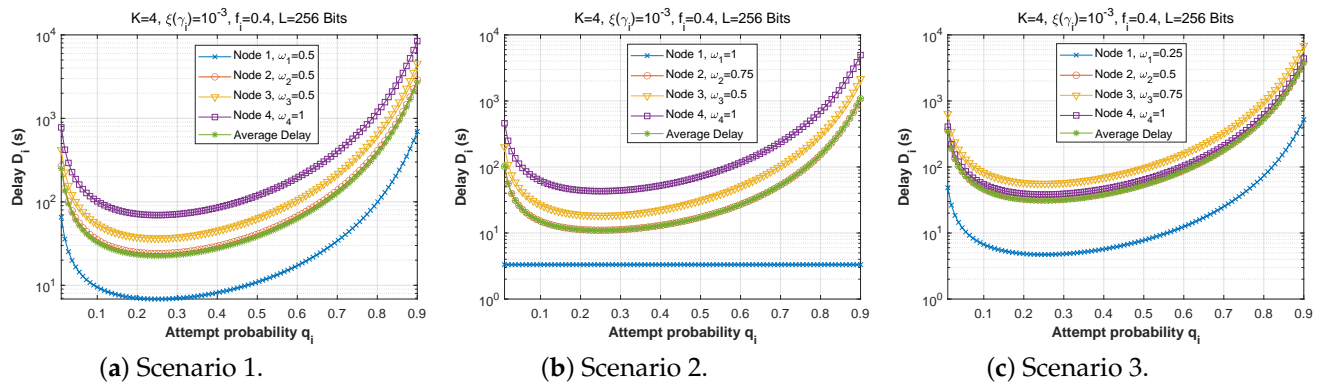
(**a**) Scenario 1.   (**b**) Scenario 2.   (**c**) Scenario 3.

Figure 15. The delay experienced at each mobile device when varying the attempt probability with the arrival rate λ_i^Q.

5.1.3. Fraction of Cellular Traffic ω_i

Figure 16 demonstrates how delay changes as a function of the fraction of cellular traffic in the first scenario when the bit error rate varies. We can see that the higher the proportion of cellular traffic, the lower the delay for a good channel. Node 4 (the nearest to the base station) always uses a value of $w_4 = 1$, as it only has one option (transmit the packets directly to the base station). The delay is plotted as a function of the fraction of cellular traffic for various regimes in Figure 17. The three subfigures are practically identical, despite the change in the arrival flow in the queue.
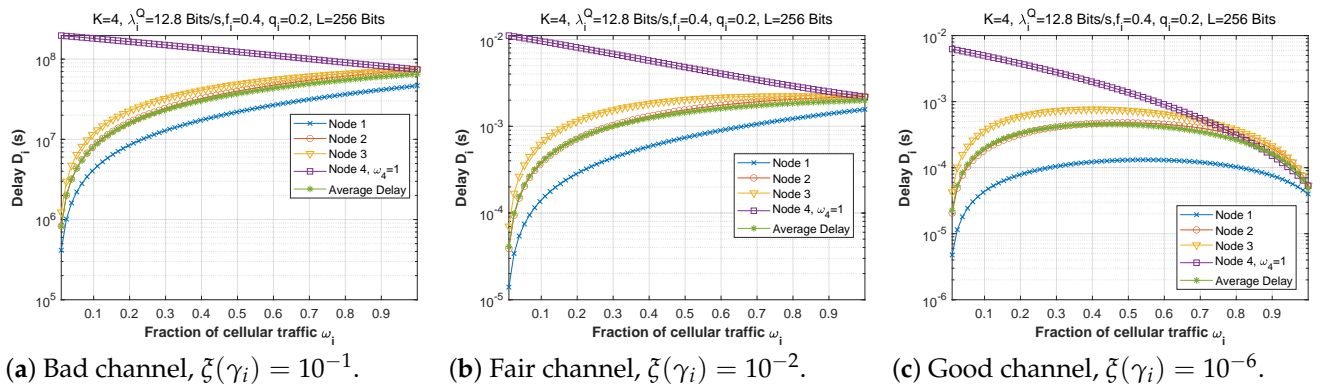

(**a**) Bad channel, $\xi(\gamma_i) = 10^{-1}$.   (**b**) Fair channel, $\xi(\gamma_i) = 10^{-2}$.   (**c**) Good channel, $\xi(\gamma_i) = 10^{-6}$.

Figure 16. Setting 1: The delay experienced at each mobile device when varying the fraction of cellular traffic with the bit error rate $\xi(\gamma_i)$.
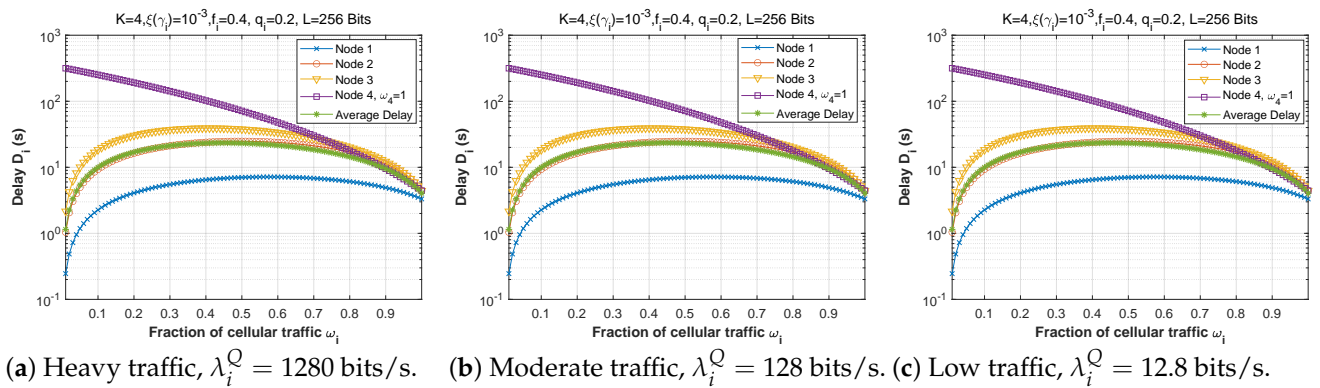

(**a**) Heavy traffic, $\lambda_i^Q = 1280$ bits/s.   (**b**) Moderate traffic, $\lambda_i^Q = 128$ bits/s. (**c**) Low traffic, $\lambda_i^Q = 12.8$ bits/s.

Figure 17. Setting 1: The delay experienced at each mobile device when varying the fraction of cellular traffic with the arrival rate λ_i^Q.

5.1.4. Arrival Rate in OWN queue λ_i^Q

We turn now to plot the delay versus the arrival rate for the three scenarios (see Figure 18). For the good channel case, it is shown that the delay increases with an increasing arrival rate for all scenarios. It is apparent that the first node is almost stable, this is because it only carried its own packets.
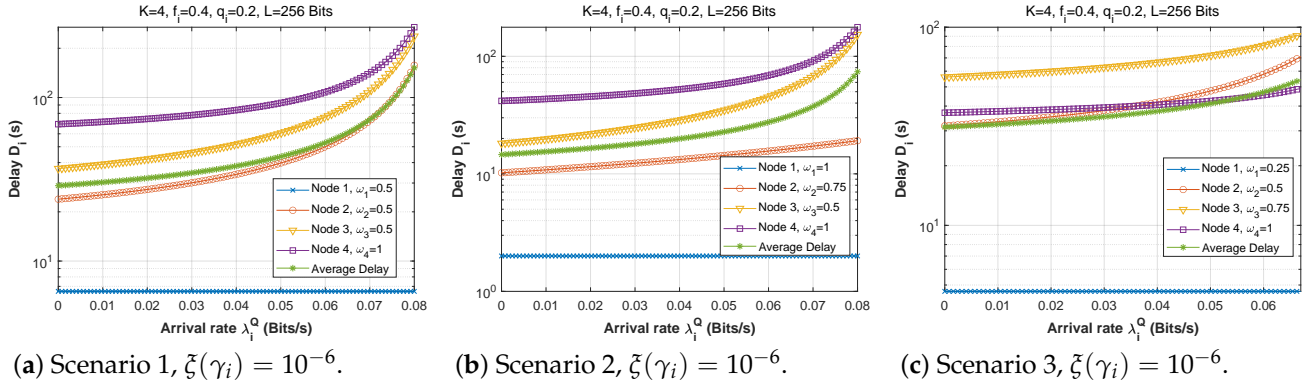

(**a**) Scenario 1, $\xi(\gamma_i) = 10^{-6}$. (**b**) Scenario 2, $\xi(\gamma_i) = 10^{-6}$. (**c**) Scenario 3, $\xi(\gamma_i) = 10^{-6}$.

Figure 18. The delay experienced at each mobile device when varying the arrival rate λ_i^Q.

5.2. AoI Simulation

This section examines the behavior of the AoI metric when the fundamental parameters $(f_i, q_i, \omega_i, \lambda_i^Q)$ change.

5.2.1. Forwarding Probability f_i

Figures 19–21 depict the AoI as a function of forwarding probability for three values of the bit error rate, and for scenarios 1, 2, and 3. It is obvious that when the forwarding probability is too high, the system suffers from a high AoI and, thus, the AoI per node is high, explaining that an arriving packet cannot be forwarded immediately, due to both a busy MAC layer as well as other packets having priority in the queue. Moreover, passing through multiple nodes (ad hoc network) automatically implies a high AoI.
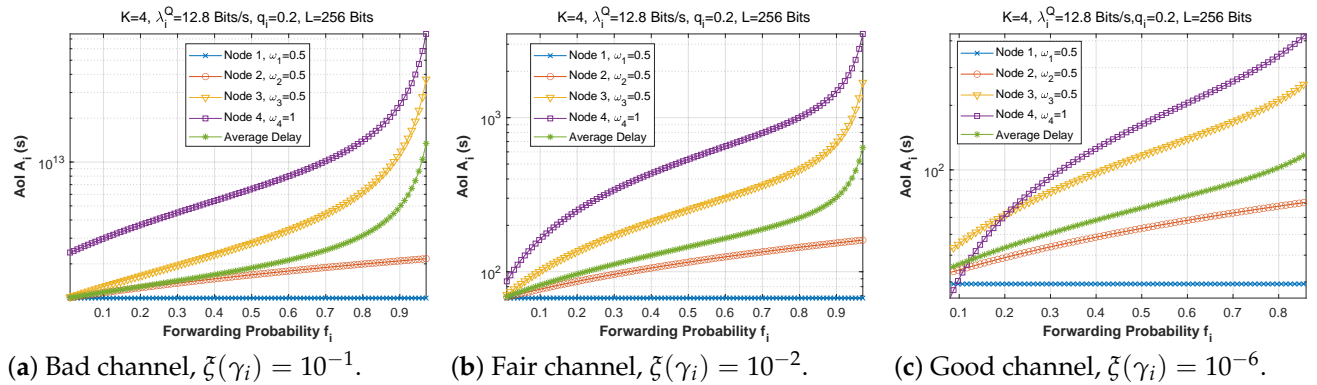

(**a**) Bad channel, $\xi(\gamma_i) = 10^{-1}$. (**b**) Fair channel, $\xi(\gamma_i) = 10^{-2}$. (**c**) Good channel, $\xi(\gamma_i) = 10^{-6}$.

Figure 19. Setting 1: The AoI experienced at each mobile device when varying the forwarding probability with the bit error rate $\xi(\gamma_i)$.
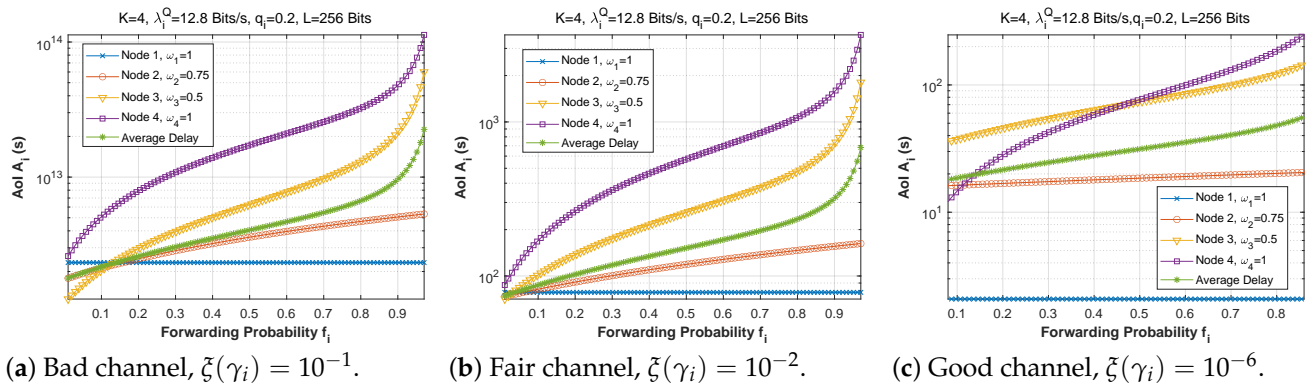
(a) Bad channel, $\xi(\gamma_i) = 10^{-1}$. **(b)** Fair channel, $\xi(\gamma_i) = 10^{-2}$. **(c)** Good channel, $\xi(\gamma_i) = 10^{-6}$.

Figure 20. Setting 2: The AoI experienced at each mobile device when varying the forwarding probability with the bit error rate $\xi(\gamma_i)$.
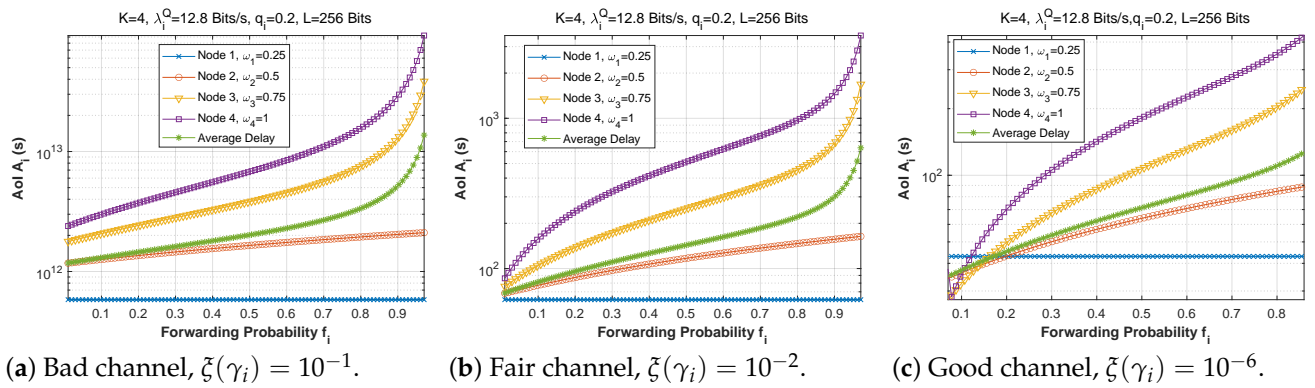

(a) Bad channel, $\xi(\gamma_i) = 10^{-1}$. **(b)** Fair channel, $\xi(\gamma_i) = 10^{-2}$. **(c)** Good channel, $\xi(\gamma_i) = 10^{-6}$.

Figure 21. Setting 3: The AoI experienced at each mobile device when varying the forwarding probability with the bit error rate $\xi(\gamma_i)$.

Figures 22–24 depict the AoI as a function of forwarding probability for all three scenarios when the arrival rate in the queue is varied. As observed before, the AoI rises with forwarding probability, and rises all the more quickly the further upstream is the node in the chain (while node 1 has a constant AoI regardless of f_i). It is noteworthy that the load on the Q queue load has little influence on AoI, while the F queue load has a significant impact.

(a) Heavy traffic, $\lambda_i^Q = 1280$ bits/s. **(b)** Moderate traffic, $\lambda_i^Q = 128$ bits/s. **(c)** Low traffic, $\lambda_i^Q = 12.8$ bits/s.
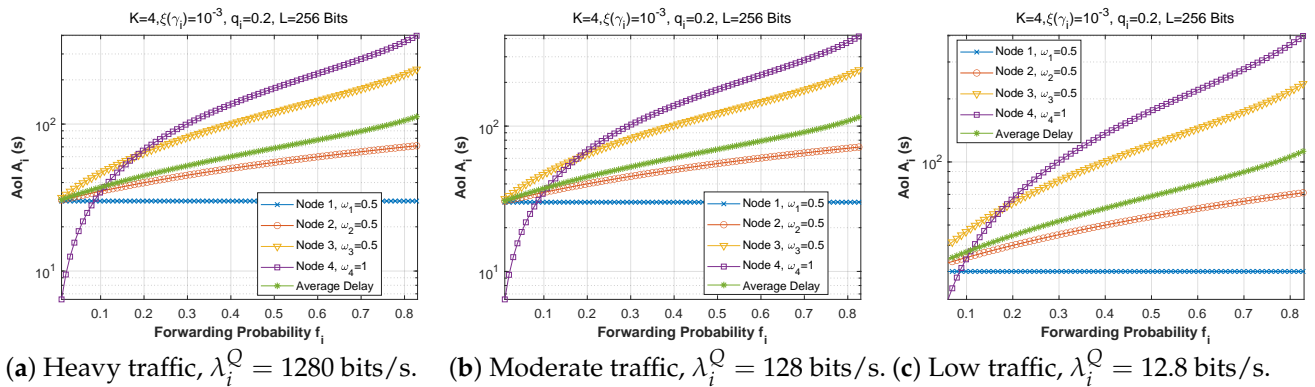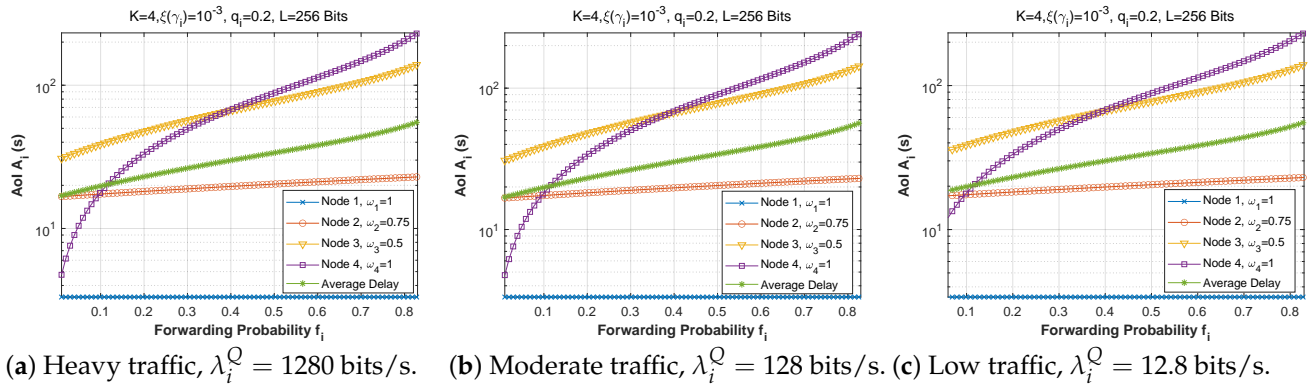
Figure 22. Setting 1: The AoI experienced at each mobile device as a function of forwarding probability f_i for various values of the arrival rate λ_i^Q.

(**a**) Heavy traffic, $\lambda_i^Q = 1280$ bits/s. (**b**) Moderate traffic, $\lambda_i^Q = 128$ bits/s. (**c**) Low traffic, $\lambda_i^Q = 12.8$ bits/s.
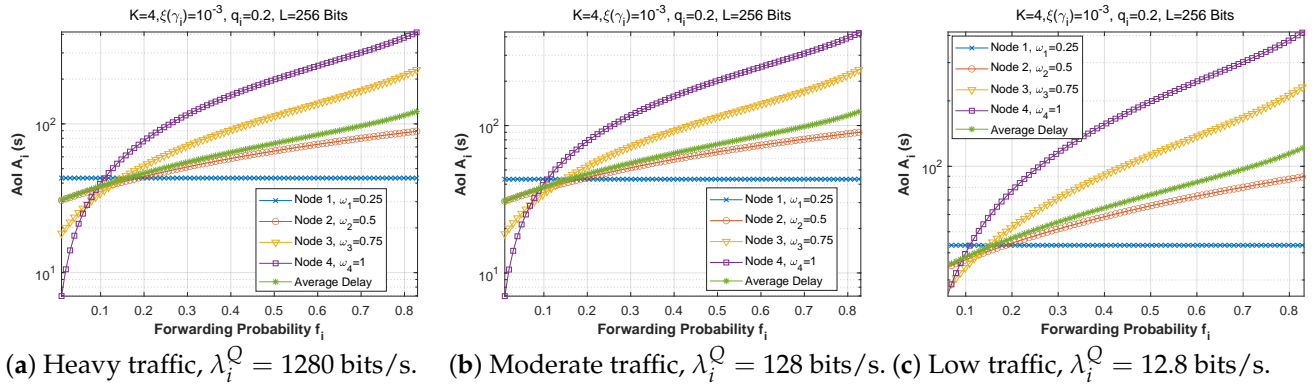
Figure 23. Setting 2: The AoI experienced at each mobile device as a function of forwarding probability f_i for various values of the arrival rate λ_i^Q.

(**a**) Heavy traffic, $\lambda_i^Q = 1280$ bits/s. (**b**) Moderate traffic, $\lambda_i^Q = 128$ bits/s. (**c**) Low traffic, $\lambda_i^Q = 12.8$ bits/s.

Figure 24. Setting 3: The AoI experienced at each mobile device as a function of forwarding probability f_i for various values of the arrival rate λ_i^Q.

5.2.2. Attempt Probability q_i

Figures 25–27 show the effect of attempt probability on AoI when the bit error rate is adjusted for the three proposed scenarios. When the attempt probability is too low, the system becomes unstable, and the AoI begins to decrease as the attempt probability decreases, reaching a minimum value when q_i is between 0.2 and 0.4. As the attempt probability increases, the queues become more congested, and the AoI increases. Because of the greater attempt probability, more packets will compete for transmission via the ad hoc network. This will tend to overload forwarding queues. Thus, the node farthest from the base station has the largest AoI compared to the node closest to the base station. Since packets in the farthest nodes looking to reach the base station as the target must spend time waiting in each node along the path, and for scenario 2, the first node with $W_1 = 1$ retains a stable AoI.
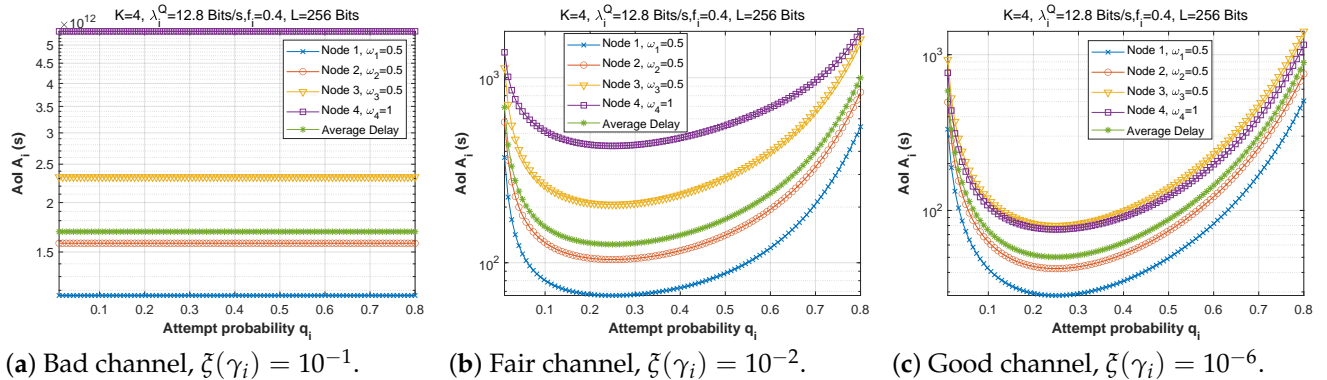

(**a**) Bad channel, $\xi(\gamma_i) = 10^{-1}$. (**b**) Fair channel, $\xi(\gamma_i) = 10^{-2}$. (**c**) Good channel, $\xi(\gamma_i) = 10^{-6}$.

Figure 25. Setting 1: The AoI experienced at each mobile device as a function of attempt probability for various bit error rates $\xi(\gamma_i)$.
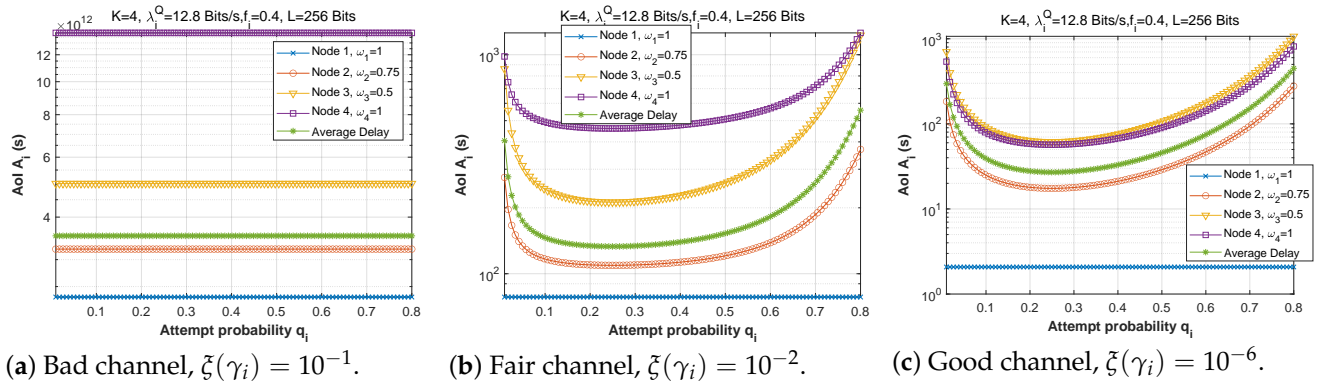
(**a**) Bad channel, $\xi(\gamma_i) = 10^{-1}$. (**b**) Fair channel, $\xi(\gamma_i) = 10^{-2}$. (**c**) Good channel, $\xi(\gamma_i) = 10^{-6}$.

Figure 26. Setting 2: The AoI experienced at each mobile device as a function of attempt probability for various bit error rates $\xi(\gamma_i)$.

(**a**) Bad channel, $\xi(\gamma_i) = 10^{-1}$. (**b**) Fair channel, $\xi(\gamma_i) = 10^{-2}$. (**c**) Good channel, $\xi(\gamma_i) = 10^{-6}$.
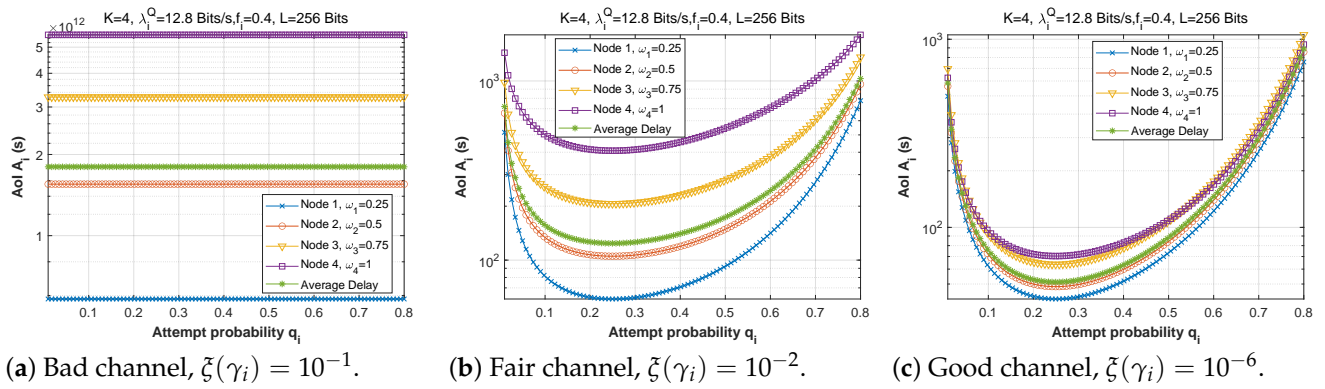
Figure 27. Scenario 3: The AoI experienced at each mobile device as a function of attempt probability for various bit error rates $\xi(\gamma_i)$.

Next, Figures 28 and 29 plot the AoI as a function of attempt probability for various traffic regimes in all three scenarios. Again, it can be observed that the arrival rate in the queue has little impact on the AoI. For scenarios 1 and 3, the subfigures are practically identical, despite the change in the arrival flow in the queue.
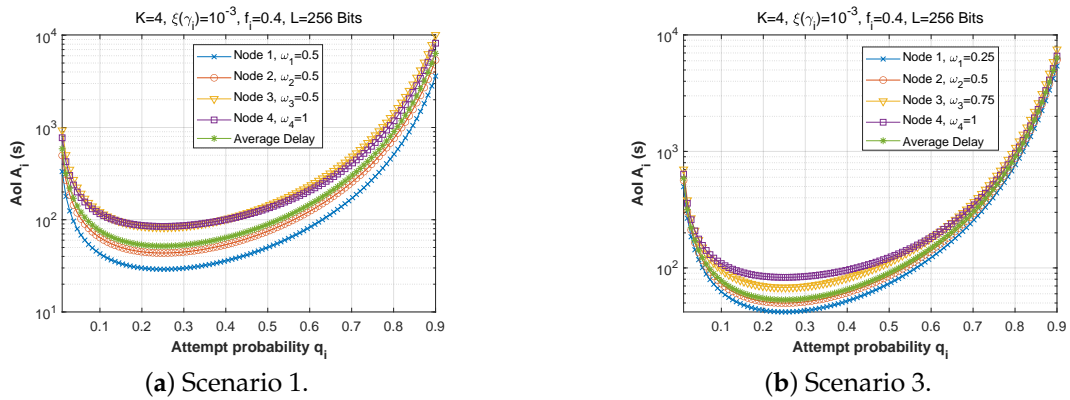

(**a**) Scenario 1. (**b**) Scenario 3.

Figure 28. The AoI experienced at each mobile device as a function of attempt probability q_i for various arrival rates λ_i^Q.

(**a**) Heavy traffic, $\lambda_i^Q = 1280$ bits/s. (**b**) Moderate traffic, $\lambda_i^Q = 128$ bits/s. (**c**) Low traffic, $\lambda_i^Q = 12.8$ bits/s.
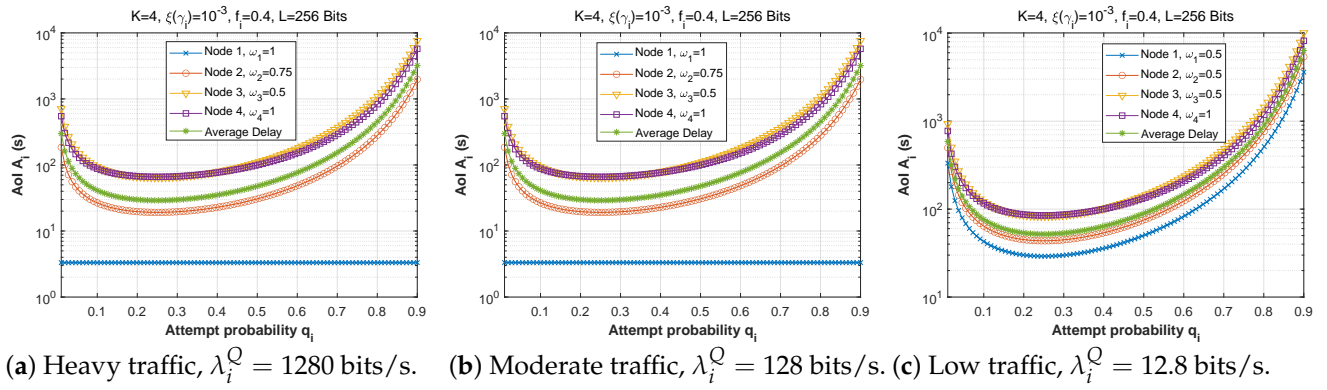
Figure 29. Setting 2: The AoI experienced at each mobile device as a function of attempt probability q_i for various arrival rates λ_i^Q.

5.2.3. Fraction of Cellular Traffic ω_i

For the first scenario, the AoI is plotted as a function of the fraction of cellular traffic in Figure 30. We demonstrate that for a good channel as the fraction of cellular traffic rises, the AoI decreases significantly until it achieves a minimum for $\omega_i = 0.9$. This decrease in AoI is justified by the fact that nodes send data directly to the base station, implying that packets arrive at their destination faster. Furthermore, the forwarding probability for nodes 1, 2, and 3 is 0.5, resulting in a congested forwarding queue at node 4, explaining its position as the node with the highest AoI.
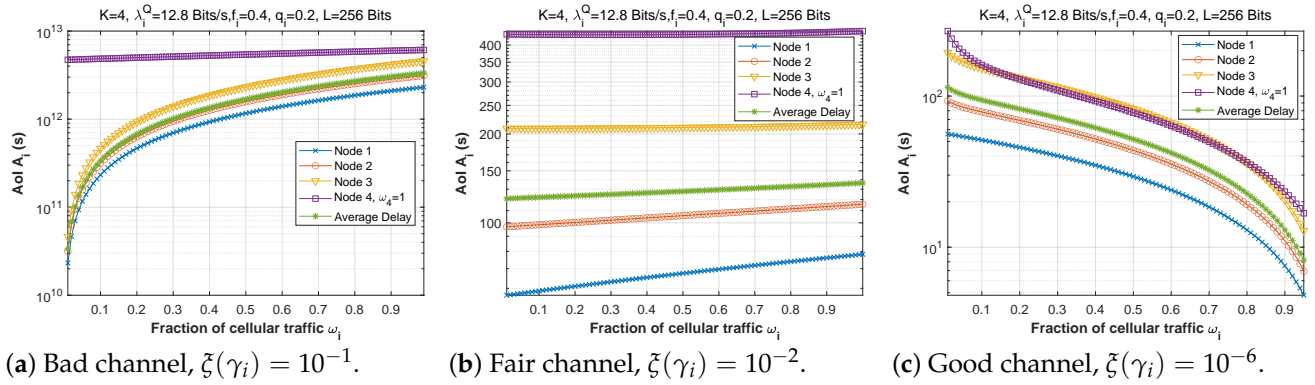

(**a**) Bad channel, $\xi(\gamma_i) = 10^{-1}$. (**b**) Fair channel, $\xi(\gamma_i) = 10^{-2}$. (**c**) Good channel, $\xi(\gamma_i) = 10^{-6}$.

Figure 30. Setting 1: The AoI experienced at each mobile device when varying the fraction of cellular traffic with the bit error rate $\xi(\gamma_i)$.

Next, the AoI is plotted as a function of the fraction of cellular traffic for different values of λ_i^Q in Figure 31. It appears that the system maintains the same behavior regardless of the rate of arrival in the queue Q.
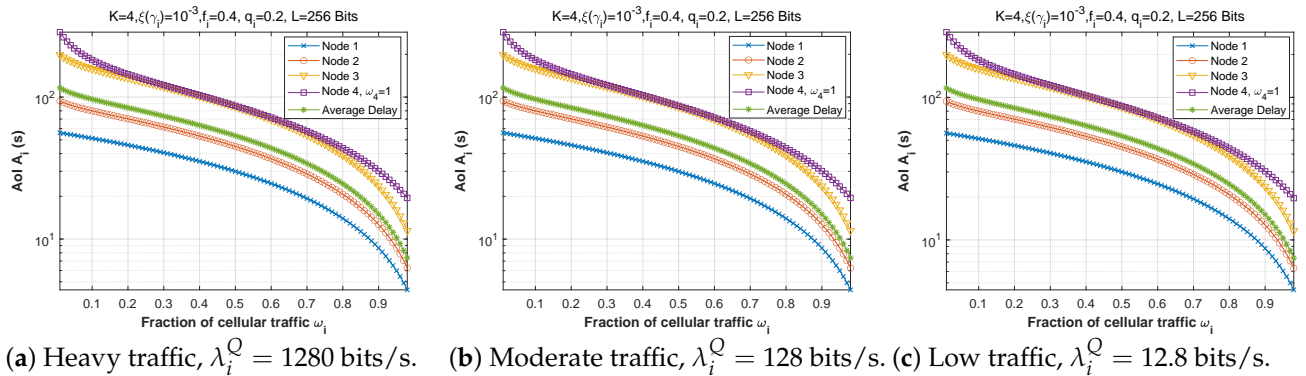

(**a**) Heavy traffic, $\lambda_i^Q = 1280$ bits/s. (**b**) Moderate traffic, $\lambda_i^Q = 128$ bits/s. (**c**) Low traffic, $\lambda_i^Q = 12.8$ bits/s.

Figure 31. Setting: The AoI experienced at each mobile device when varying the fraction of cellular traffic with the arrival rate λ_i^Q.

5.2.4. Arrival Rate in its Own Queue λ_i^Q

Finally, the AoI is plotted as a function of the arrival rate in queue Q, for all three scenarios in Figure 32. In all cases, the AoI is extremely high at low arrival rates. This is because such low levels of traffic imply insufficient status updates at the base station. As λ_i^Q is allowed to increase, the AoI then drops at all nodes until it reaches a minimum value, then rises again, as the queues start to fill up and the system moves toward saturation. The AoI also increases according to the distance of a node from the base station, as this relates to the number of required hops to reach it.
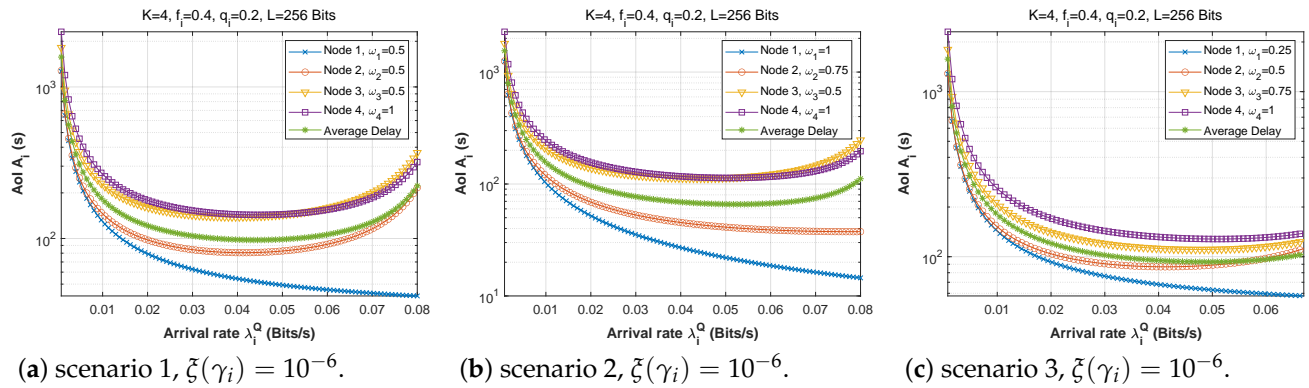

(**a**) scenario 1, $\xi(\gamma_i) = 10^{-6}$.　　(**b**) scenario 2, $\xi(\gamma_i) = 10^{-6}$.　　(**c**) scenario 3, $\xi(\gamma_i) = 10^{-6}$.

Figure 32. The AoI experienced at each mobile device when varying the arrival rate λ_i^Q.

6. Concluding Remarks

Multi-hop networks promise to efficiently collect data from IoT devices deployed in a target area as well as relay their data to legacy systems, such as cellular networks. In this paper, we propose a comprehensive theoretical framework for analyzing and understanding the dynamics of such a network. The suggested model is intended to assist mostly in the planning and sizing of an IoT network, as a means of ensuring target/satisfactory performance and effective deployment. We provide a queuing–theory-based model that allows for cross-layered optimization across the APP, NET, MAC, and PHY layers. The suggested model was evaluated using a discrete-event simulation, and it accurately predicts network performance. Our model can measure E2E delay and AoI, making it an excellent choice for evaluating the freshness of information for active streams. It is necessary to examine the impact of forwarding probability, attempt probability, a fraction of cellular traffic, the arrival rate in the queue, and other parameters. The determination of the stability region as a function of these factors constitutes an end result of interest. Many trade-offs have been outlined, as well as a thorough discussion of parameter tuning and network design. This article opens up the way for many exciting areas, including network design and optimal configuration, energy efficiency, wireless energy transfer, flexible infrastructure, etc. Future extensions of this work will examine energy consumption measures to adjust network parameters in order to ensure limited and/or balanced energy consumption.

Author Contributions: Conceptualization, I.C., E.S.; Methodology, I.C. and E.S.; Validation, E.S., R.A. and S.R.; Formal analysis, I.C., E.S. and S.R.; Investigation, I.C. and E.S.; Writing – review and editing, I.C.; Supervision, E.S., R.A., S.R. and M.S.; Project administration, E.S. All authors have read an agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

References

1. Pollakis, E.; Cavalcante, R.L.G.; Stanczak, S. Enhancing energy efficient network operation in multi-RAT cellular environments through sparse optimization. In Proceedings of the 2013 IEEE 14th Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Darmstadt, Germany, 16–19 June 2013; pp. 260–264.
2. Rault, T.; Bouabdallah, A.; Challal, Y. Energy efficiency in wireless sensor networks: A top-down survey. *Comput. Netw.* **2014**, *67*, 104–122.
3. Kim, J.; Lee, H.W.; Chong, S. Super-MAC Design for Tightly Coupled Multi-RAT Networks. *IEEE Trans. Commun.* **2019**, *67*, 6939–6951.
4. El-Azouzi, R.; Sabir, E.; Samanta, S.K.; El-Khoury, R.; Bouyakhf, E.H. An end-to-end QoS framework for IEEE 802.16 and ad hoc integrated networks. In Proceedings of the 6th International Conference on Mobile Technology, Application & Systems, Nice, France, 2–4 September 2009; pp. 1–8.
5. Wen, J.; Sheng, M.; Wang, X.; Li, J.; Sun, H. On the capacity of downlink multi-hop heterogeneous cellular networks. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 4092–4103.
6. Chen, H.; Gu, Y.; Liew, S.C. Age-of-information dependent random access for massive IoT networks. *arXiv* **2020**, arXiv:2001.04780.
7. Zhou, B.; Saad, W. On the Age of Information in Internet of Things Systems with Correlated Devices. *arXiv* **2020**, arXiv:2001.11162.
8. Han, S. Congestion-aware WiFi offload algorithm for 5G heterogeneous wireless networks. *Comput. Commun.* **2020**, *164*, 69–76.
9. Lim, G.; Xiong, C.; Cimini, L.J.; Li, G.Y. Energy-efficient resource allocation for OFDMA-based multi-RAT networks. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 2696–2705.
10. Fadel, M.; Ibrahim, A.S.; Elgebaly, H. Qos-aware multi-rat resource allocation with minimum transmit power in multiuser ofdm system. In Proceedings of the 2012 IEEE Globecom Workshops, Anaheim, CA, USA, 3–7 December 2012; pp. 670–675.
11. Zhang, W.; Duan, D.; Yang, L. Relay selection from a battery energy efficiency perspective. In Proceedings of the MILCOM 2009-2009 IEEE Military Communications Conference, Boston, MA, USA, 18–21 October 2009; pp. 1–7.
12. Lim, G.; Cimini, L.J. Energy-efficient cooperative relaying in heterogeneous radio access networks. *IEEE Wirel. Commun. Lett.* **2012**, *1*, 476–479.
13. Carvalho, G.H.; Woungang, I.; Anpalagan, A.; Hossain, E. QoS-aware energy-efficient joint radio resource management in multi-RAT heterogeneous networks. *IEEE Trans. Veh. Technol.* **2015**, *65*, 6343–6365.
14. Kherani, A.; El-Khoury, R.; El-Azouzi, R.; Altman, E. Stability–throughput tradeoff and routing in multi-hop wireless ad hoc networks. *Comput. Netw.* **2008**, *52*, 1365–1389.
15. El-Azouzi, R.; Sabir, E.; Raiss-El-Fenni, M.; Samanta, S.K. A Unified NET-MAC-PHY cross-layer framework for performance evaluation of multi-hop Ad hoc WLANs. *EAI Endorsed Trans. Mob. Commun. Appl.* **2014**, *1* .
16. El-Khoury, R.; El-Azouzi, R.; Altman, E. Delay analysis for real-time streaming media in multi-hop ad hoc networks. In Proceedings of the 2008 6th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks and Workshops, Berlin, Germany, 1–3 April 2008; pp. 419–428.
17. Ayan, O.; Gürsu, H.M.; Papa, A.; Kellerer, W. Probability analysis of age of information in multi-hop networks. *IEEE Netw. Lett.* **2020**, *2*, 76–80.
18. Talak, R.; Karaman, S.; Modiano, E. Minimizing age-of-information in multi-hop wireless networks. In Proceedings of the 2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 3–6 October 2017; pp. 486–493.
19. Kaul, S.K.; Yates, R.D.; Gruteser, M. Status updates through queues. In Proceedings of the 2012 46th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, 21–23 March 2012; pp. 1–6.
20. Bedewy, A.M.; Sun, Y.; Shroff, N.B. Age-optimal information updates in multihop networks. In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–27 June 2017; pp. 576–580.
21. Yates, R.D.; Sun, Y.; Brown, D.R.; Kaul, S.K.; Modiano, E.; Ulukus, S. Age of information: An introduction and survey. *IEEE J. Sel. Areas Commun.* **2021**.
22. Chen, C.; Ma, J.; Yu, K. Designing energy-efficient wireless sensor networks with mobile sinks. In Proceedings of the 4th ACM Conference on Embedded Networked Sensor Systems (SenSys 2006), Boulder, CO, USA, 31 Octoer–3 November 2006.
23. Kam, C.; Kompella, S.; Nguyen, G.D.; Wieselthier, J.E.; Ephremides, A. On the age of information with packet deadlines. *IEEE Trans. Inf. Theory* **2018**, *64*, 6419–6428.
24. Hu, Y.; Zhong, Y.; Zhang, W. Age of information in Poisson networks. In Proceedings of the 2018 10th International Conference on Wireless Communications and Signal Processing (WCSP), Hangzhou, China, 18–20 October 2018; pp. 1–6.
25. Althoubi, A.; Alshahrani, R.; Peyravi, H. Delay analysis in iot sensor networks. *Sensors* **2021**, *21*, 3876.
26. Abdelradi, Y.M.; El-Sherif, A.A.; Afify, L.H. A queueing theory approach to traffic offloading in heterogeneous cellular networks. *AEU-Int. J. Electron. Commun.* **2021**, *139*, 153910.

27. Chinchilla-Romero, L.; Prados-Garzon, J.; Ameigeiras, P.; Muñoz, P.; Lopez-Soler, J.M. 5G Infrastructure Network Slicing: E2E Mean Delay Model and Effectiveness Assessment to Reduce Downtimes in Industry 4.0. *Sensors* **2021**, *22*, 229.

28. Hasan, M.Z.; Al-Turjman, F.; Al-Rizzo, H. Analysis of cross-layer design of quality-of-service forward geographic wireless sensor network routing strategies in green internet of things. *IEEE Access* **2018**, *6*, 20371–20389.

29. Blaszczyszyn, B.; Muhlethaler, P.; Banaouas, S. *A comparison of ALOHA and CSMA in Wireless ad Hoc Networks under Different Channel Conditions*; HAL: Lyon, France, 2010; INRIA-00530093.

30. Saad, W.; Han, Z.; Zheng, R.; Debbah, M.; Poor, H.V. A college admissions game for uplink user association in wireless small cell networks. In Proceedings of the IEEE INFOCOM 2014-IEEE Conference on Computer Communications, Toronto, ON, Canada, 27 April–2 May 2014; pp. 1096–1104.

31. Habbal, A.; Goudar, S.I.; Hassan, S. Context-aware radio access technology selection in 5G ultra dense networks. *IEEE Access* **2017**, *5*, 6636–6648.

32. Huang, L.; Modiano, E. Optimizing age-of-information in a multi-class queueing system. In Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, China, 14–19 June 2015; pp. 1681–1685.